DOCUMENT RESUME

ED 242 781 TM 840 199

AUTHOR	Nitko, Anthony J.; Hsu, Tse-chi
TITLE	Item Analysis Appropriate for Domain-Referenced Classroom Testing. (Project Technical Report Number 1).
INSTITUTION	Pittsburgh Univ., Pa.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	Apr 84
CONTRACT	400-82-0014
NOTE	146p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
AVAILABLE FROM	Chairman, Department of Educational Research Methodology, 5C03 Forbes Quadrangle, University of Pittsburgh, Pittsburgh, PA 15260 ($6.00)
PUB TYPE	Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	Computer Assisted Testing; Criterion Referenced Tests; *Item Analysis; Microcomputers; Sampling; *Statistical Analysis; *Test Construction; Test Items; Test Use

ABSTRACT

Item analysis procedures appropriate for domain-referenced classroom testing are described. A conceptual framework within which item statistics can be considered and promising statistics in light of this framework are presented. The sampling fluctuations of the more promising item statistics for sample sizes comparable to the typical classroom size are described. Several statistical indices are recommended for use in an item analysis package programmed for an Apple II Plus microcomputer. The primary purposes of an item analysis of classroom tests are to inform the teacher about strengths and weaknesses of the class in relation to skills measured by test items, as well as to determine which items need to be replaced or revised. A secondary purpose of item analysis is selection of items from an item bank to improve the utility of a test for specific purposes. Several statistical indices are reviewed and classified as basic, recommended, or not recommended. (DWH)

ITEM ANALYSIS APPROPRIATE FOR DOMAIN-REFERENCED CLASSROOM TESTING

by

Anthony J. Nitko
and
Tse-chi Hsu

University of Pittsburgh

(Project Technical Report Number 1)

APRIL 1983, AERA APRIL 1984

The work upon which this publication is based was performed pursuant to Contract No. 400-82-0014 of the National Institute of Education. It does not, however, necessarily reflect the views of that agency.

Appropriate Microcomputer Item Analysis

for Domain-Referenced Classroom Testing

Anthony J. Nitko and Tse-chi Hsu

School of Education

University of Pittsburgh

This paper describes item analysis procedures appropriate for domain-referenced classroom testing and how these procedures can be implemented with a microcomputer program. First, it presents a conceptual framework within which teachers' informational needs and item statistics can be considered. Second, we review approximately fifty item statistics, using logical analysis and Monte Carlo sampling studies to ultimately recommend several statistics to be incorporated into a microcomputer program for classroom teachers.

TABLE OF CONTENTS

Page

Item Analysis Appropriate for Domain-Referenced Classroom Testing*

by
Anthony J. Nitko and Tse-chi Hsu
School of Education
University of Pittsburgh

The purpose of this paper is to describe the kinds of item analysis information useful for domain-referenced classroom testing. The paper is organized in the following way. First we present a conceptual framework within which item statistics can be considered. Second, we review promising statistics in light of this framework. Third, we examine the sampling fluctuations of several of the more promising item statistics for sample sizes comparable to what we would expect the typical classroom size to be. Fourth, we recommend several statistical indices that are the most promising ones to use in an item analysis package programmed for an Apple II Plus microcomputer.

The reader of this report should keep in mind several points. First, the item analysis procedures and statistics recommended in this report are constrained by the practical limits of schools, of teachers' experience and time, and of the capacities of a particular microcomputer. Second, the primary functions of an analysis of pupils' responses to test items are to assist a teacher in (a) making instructionally relevant decisions and (b) improving the technical quality of the test items used. In this item analysis process, the teacher is encouraged to use the computer as a tool and no

*We are deeply indebted to Dr. Huynh Huynh for his valuable assistance in providing us with the Monte Carlo sampling data presented in the third section and for his immense contribution in clarifying our thinking about the item statistics reviewed in this report. We are solely responsible, of course, for all errors and misconceptions that remain herein.

attempt is made to use item statistics to create a computer-assisted "teacher proof" system of item analysis. Third, it is recognized that selecting appropriate item statistics means not simply focusing on the quality and usefulness of the statistics qua statistics, but also means considering the appropriateness of the statistics in terms of the understanding and interpretation that teachers are able to give them. Fourth, the appropriate number and the presentation of statistics is important to their use by teachers. If too many statistical indices are provided simultaneously and in an "unfriendly" format, a teacher will be confused. Thus, although we recommend quite a few statistical indices in this report, we do not recommend that these statistics be reported simultaneously in uninterpreted form. Designing an item analysis microcomputer program is in part a human engineering problem. Fifth, a microcomputer program that computes the recommended statistics should present the information to the teacher in a way that will facilitate interpreting the teacher's particular classroom data. Sometimes this means simply displaying the numerical value of a statistical index. At other times it will mean programming decision rules into the computer that will recommend certain teacher actions or certain teacher options. Sixth, it should be noted that all of the statistical indices we recommend in the last section should be available to a teacher upon request, even if they are not displayed initially. Thus programming techniques should be used that will permit a teacher to dip deeper into the data and to obtain the actual numerical values of the indices, if desired.

Frameword for Considering Item Analysis and Item Statistics

There are important differences between using tests to measure pupils and using tests to improve the pupils' instruction: Whereas measurement seeks not to alter the characteristic being tested, instruction explicitly seeks to change the pupil so that eventually every test item in the domain can be answered correctly (cf. Lord, 1970). In order for tests to be effective as classroom instructional tools, however, it is necessary to integrate them into the instructional decision-making process. This means that teachers have to design tests for the decisions for which they will be using them. This assures that the test information has a reasonable chance of being useful.

The term domain-referenced test is broadly defined to mean a test that is built so that scores on it can be referenced to a well-defined class or domain of behaviors in a way that permits an examinee's status on that domain to be estimated. This is a broad definition of domain-referencing and there is little difference between it and criterion-referencing as this latter term has been recently explicated (Nitko, 1980). Both concepts essentially mean the same thing, requiring a well-defined class of tasks or behaviors to which test performance can be referenced. Most persons prefer the term criterion-referencing (Popham, 1978; Hambleton, Swaminathan, Algina, & Coulson, 1978).

Classifications of domain-referenced tests such as that presented by Nitko (1980) are likely to be unfamiliar to teachers. However, teachers can be encouraged to view their own tests in this broader context. Most teachers' tests are of the unordered variety, being built on the basis of verbal statements of stimuli and responses (i.e., behavioral objectives) and sometimes on the basis of diagnostic categories of pupil difficulties. But at least for some classroom decisions, such as grouping students or distinguishing among degrees of mastery of a topic, ordered domains may be more appropriate. This means, for example, that a teacher's interpretation of item analysis and other test statistics will depend on the type of domain-referenced test being built, as well as on the type of decision for which the test information will be used.

The workable level of specificity for domain definition is likely to be the behavioral objective. Teachers can use behavioral objectives to orgarize and direct their instruction. Currently many training programs teach teachers how to write objectives and use them for instructional design. Further, many school districts define their curriculum using objectives. Thus, for most teachers, domain-referenced classroom testing is likely to center around behavioral objectives at this point in time.

The responses of students to the items on a classroom test provide a teacher with information in three broad and interrelated areas: (1) improving and guiding instruction, (2) editing and improving individual test items, and (3) improving the properties of the total test score for certain decision-making purposes. Pupils' responses to stimulus material a teacher presents for purposes of evaluation provide clues concerning what pupils have learned, the extent to which the material has been learned, and the nature of pupils' errors and misunderstandings. Item analysis can provide a teacher with valuable summary information about the class of pupils, as well as identify pupils who respond in unusual ways to the stimulus material. Such instructionally relevant information when brought to the attention of a teacher can provide the basis for instructional planning.

Second, pupils' responses to test items provide valuable information about how the individual test items are functioning. Test items should be designed to elicit certain important pupil responses that a teacher can use to decide whether learning has occurred. Viewed in this way, a test item and its parts have very specific functions. Data about the test item and its parts can be analyzed and used to decide whether these functions are being fulfilled. As an example, consider the alternatives of a multiple-choice item. Data can be gathered to provide a teacher with information about such matters as whether less knowledgeable students are attracted to incorrect responses and whether two or more alternatives appear to be ambiguous to the more knowledgeable pupils. Additional information about how an item has functioned and what might be done to improve it can be provided, of course.

A third area in which item statistics can be helpful is in suggesting ways for improving the entire collection or ensemble of test items that comprise a particular test. Each item contributes to the score on the total test in well known ways. Thus, the entire test is dependent on the properties of the individual items. What is considered to be the desirable properties of the total test, on the other hand, depends on the particular purposes or decisions for which that test score will be used. A test may be used, for example, to estimate a domain score without reference to the performance of other pupils in the class. Or, the test score may form the basis to rank or order pupils for purposes of assigning letter grades or for forming subgroupings of pupils for instructional purposes. Tests with such diverse purposes will have different properties. The items comprising the tests will need to exhibit different properties as well. Thus, statistics computed in an item analysis microcomputer program will need to fit into the purposes for which the teacher will use the total test scores.

The three broad areas and the specific kinds of information needed under each area are listed in Table 1. The specific information is discussed in the sections which follow. As can be seen from a perusal of the table, the three areas are interrelated and specific information in one area may often be used for a related purpose in another area. In the discussion which follows, each kind of specific information is discussed separately. However, the reader should keep in mind their interrelationships.

Insert Table 1 here

Item Analysis Information Useful for Guiding Instruction

Unless otherwise noted, the descriptions in this section refer to information that is provided for each test item, rather than for the total test or for clusters of test items.

1. Summary of how the class performed. Each test item is intended to measure knowledge or application of an important fact, concept, or principle that a teacher has taught. Often, such knowledge and/or application is prerequisite to the next unit or step in an instructional sequence. It is useful, therefore, for a teacher to know the extent to which the class as a whole has acquired this knowledge or skill since future instructional planning can be informed by such information.

2. Discrepancy between a teacher's expectation of a class' performance and the actual performance of the class. An important kind of information for a teacher is whether the class performed as the teacher expected. To conduct instruction effectively a teacher needs a good sense of whether students are behaving in expected ways. Teachers often do have informal, implicit expectations about the number or percent of students who would be expected to have learned certain concepts at particular points in time. An item analysis program can and should compare a teacher's expectations with actual student performance and alert the teacher to confirmations or discrepancies. Knowledge and/or skill areas in which students perform significantly worse than a teacher expects can serve as the basis for planning remedial instruction, while areas in which students perform better than expected can reinforce a teacher's self-concept in relation to teaching skill and serve to raise a

Table 1. Three broad areas and the specific information needed about the items comprising a domain-referenced classroom test.

Areas in which information will be needed for each item:		
Improving and guiding instruction	Rewriting individual test items	Selecting items to put on a test
1. Summary of the performance of the class.	1. Extent of item-objective congruence.	1. Item discrimination level.
2. Discrepancy between the performance a teacher expected of the class and the actual class performance.	2. Extent of item-instructional event congruence.	2. Item difficulty level.
3. Unusual performance of a student on an item.	3. Vocabulary level of an item.	3. Relation of item to test blueprint and/or domain specification.
4. Hierarchical ordering of the items on a test.	4. Item difficulty level.	4. Estimated total test properties based on items to be included on the test.
5. Change in a class' performance after instruction.	5. Item discrimination level.	
6. Summary of the seriousness of pupils' errors on an item.	6. Identification of poor distractors.	
7. Summary of the types of errors pupils committed on an item.	7. Identification of ambiguous alternatives.	
	8. Identification of miskeyed items.	
	9. Identification of patterns of guessing among knowledgeable students.	

teacher's expectations for students in subsequent learning.

3. Unusual student performance in relation to a particular item. As students interact with test materials they can be expected to behave in a rather consistent manner. Students with less knowledge and skill can be expected to score poorly on difficult test items, but to score better on relatively easy items. Similarly, students with a good command of the subject can be expected to do well on both easy and relatively more difficult items. When a student with good ability does poorly on a relatively easy item, or when a student with poor command of knowledge or skill in an area gets a rather difficult item correct, these situations should be brought to the teacher's attention for explanation and consideration for possible action.

In like manner, test items themselves can exhibit unusual patterns. Identifying these patterns could alert the teacher to items that for some reason do not "fit-in" with the majority of items. These items may be in need of revision or the items may identify instructional areas that need attention.

4. Hierarchical ordering among the items in a particular test. It would be useful for planning instruction and diagnosis if a teacher knew whether a hierarchical structure existed among the items in a test. Depending on the nature of the skills and concepts included on the test, the identification of a hierarchy among the items could help a teacher plan diagnosis and remedial instruction by suggesting subconcept arrangements or suggesting a possible order in which concepts could be taught.

5. Changes in a class' performance as a result of instruction. Occasionally, a teacher may use a pre-instructional test (pretest) and a post-instructional test (posttest) containing identical (or equivalent)

test items. In such cases, it is important for the teacher to know the items on which the class' performance changed, and the extent and direction of that change. An item that tests a particular concept or skill and for which there is little or no change as a result of instruction or for which performance after instruction is worse than before instruction, may indicate that the instruction was ineffective or unnecessary, the item was poorly written, or the students had responded indifferently. In any event, if pretest and posttest data are available, an item analysis program should analyze it and permit the teacher who so desires to consider its implications for instruction.

6. Summary of seriousness of pupil errors on an item. Pupils who answer an item completely incorrectly or receive less than full credit on an item commit errors of various degrees of seriousness. It would be helpful to instructional planning if a teacher had, for each test item, a summary of the seriousness of errors committed. This summary could help, for example, in setting instructional priorities. Such information could be obtained, however, only if the teacher could codify or rate the seriousness of errors of each student.

7. Summary of types of pupil errors on an item. Related to Point 6 above, the type or kind of error committed is useful information as well. A teacher would need to classify the various types of errors that could be committed by students (presumably known by a teacher from past experience) and then in some way identify for each student the type(s) of error committed. This may be a tedious and, therefore, impractical task for a teacher unless (a) the number of error types is small or (b) the options on a multiple-choice test are specifically written to attract students who commit specific error types. In the former case, it is conceivable that 3 or 4 coarse types

of errors which could be found on any item on the test could be identified. For example, in social studies, errors could be classified as (a) incorrect reasoning, (b) incorrect knowledge of a concept or principle, (c) lack of knowledge of an important fact, and (d) spelling errors. Each student's response is graded in the normal fashion and, in addition, items with less than perfect responses are coded according to one or more of these error categories. (If the above illustrated error categories were arranged in order of seriousness, then both information Points 6 and 7 could be handled simultaneously.) In the case of multiple-choice items, a similar categorization could result, except that more error types could be identified because the microcomputer could automatically classify pupils' responses into various error types.

<u>Item Analysis Information Useful for Editing and Revising Items</u>

Certain kinds of information can be obtained from pupil responses to classroom test items that will suggest possible flaws in the items. Items exhibiting patterns of pupil responses suggestive of flaws could be flagged and brought to the teacher's attention. Some types of information useful for revising items may come from a closer examination of the items by the teacher, rather than analysis of pupil responses per se. Both types of information are described below.

1. <u>Extent of item-objective congruence</u>. An essential part of any review of classroom test items is the extent to which each item corresponds to the instructional objective it is intended to measure. This information can be obtained either from the judgment of an individual teacher or from the pooled judgments of a group of teachers. The former is likely to be the

typical source of information, while the latter is likely to be obtained when committees of teachers form common tests or when a school district uses an objective-based mastery learning system. In the latter situation, individual teachers usually do not construct their own test items. Often, the items are purchased or developed by outside agencies on a contract basis.

2. Congruence of test items to instructional events in a classroom. It is important that an item correspond to what a teacher has taught or the pupils were supposed to study. It often happens when test items are purchased (or provided as part of an instructional materials package) that they correspond to written statements of objectives but not to the precise manner in which students were taught to respond in the classroom. For example, a publisher-provided test item in history may emphasize a different interpretation than occurred in the classroom or a science item may illustrate a principle with a different experiement than a teacher used. In such cases, teacher revisions can "fine-tune" an item to make it a more valid measure of a pupil's learning.

3. Vocabulary appropriate to the level of the students. Items written by persons who lack daily contact with the particular pupils being tested may contain phrases and vocabulary words that are not appropriate and thereby interfere with pupils' ability to express the knowledge they have acquired. For example, in a language development curriculum in a junior high school, students may learn the definitions of new vocabulary words through class discussion and writing sentences using their current vocabulary and language level. A teacher, however, may elect to use items on a mastery test that were provided by a textbook publisher. Such items may be multiple-choice and, conceivably, their alternatives could contain vocabulary words that are beyond the language development level of the students being tested. Thus,

although students may have learned the specific words they were taught they could not demonstrate this knowledge to the teacher.

The information about the vocabulary level of the wording in an item can be obtained either by judgments from one or more teachers, or by checking an item against a specific vocabulary list.

4. Difficulty level of the item for students. An item that too few students answer correctly may be flawed in some way and hence should be revised. But difficulty level alone is not a sole criterion for revision, since a test item may be well-written but the students may not have learned the requisite material. Similarly, items that are too easy may reflect good pupil learning or an item that is too obviously correct to pupils. In either case--flawed items or reflection of the learning status of students--the difficulty level of the item contains important information.

5. The discrimination level of an item. Items for which the lower scoring pupils on a test do better than the high scoring pupils need to be examined for possible flaws, since these items function in a manner that is in opposition to the bulk of the items in the test. Similarly, items which do not distinguish the more able from the less able should be examined in at least a cursory manner to assure that they are properly written. As with all the information in this section, the purpose is to identify items that may be in need of revision, rather than to collect information for purposes of culling and selecting items.

The following information can be collected only for true-false, matching, and multiple-choice items.

6. Identification of poor distractors. The distractors of a multiple-choice item function as plausible choices for the students who have not acquired the knowledge required to answer the item correctly. Empirical data

from pupils can identify items that are not functioning in this way.

7. Identification of ambiguous alternatives. In this context, two alternatives are ambiguous if students who know the material an item is supposed to test, tend to have difficulty deciding which of the two alternatives is the correct answer.

8. Identification of miskeyed items. Occasionally a teacher inadvertantly miskeys a multiple-choice item. Data from pupil responses are examined in relation to the teacher-keyed answer. If the more knowledgeable students choose an incorrect alternative in large numbers, the items may have been miskeyed. Flagging such items bring them to the attention of the teacher.

9. Identification of items for which random guessing may be occurring among the more knowledgeable students. The more knowledgeable students are expected to have acquired the information or skill on which an item is based. Studying the response patterns of this group of students may reveal that they are not responding in the expected manner. If so, such items should be flagged and reviewed by the teacher.

Improving Properties of the Total Test*

The properties of the total test are a function of the properties of the items comprising the test. Therefore, it is important that a teacher attend to certain item properties when assembling a test. The properties of the items to which a teacher should attend depend to a considerable extent on how the total test score will be used--that is, on the decisions for which the teacher will use the scores.

In general, classroom tests tend to be used for decisions that require one of the following: (a) complete ordering of students, (b) partial ordering of students, and (c) ascertaining the domain status of students. Ranking

*We have limited this technical report to a discussion of the item statistics only although we recognized that a complete item analysis computer package should compute total test score statistics (e.g., mean, standard deviation, median, and various reliability indices), compute percentile ranks (perhaps standard scores), and tabulate a frequency distribution.

students on a test and grading on the curve are examples of test usages requiring complete ordering of students. Some uses to which test scores are put require partial ordering, as when a teacher seeks to place students into two groups--for example, better readers and less able readers--with the intention of treating individuals within each group in approximately the same way. (All students in the better readers group, for example, may be permitted to proceed with new material, while students in the lower group are given the same remedial instruction.)

A teacher seeks an estimate of a pupil's domain status when a decision depends on a person's domain score without regard to the domain scores of other pupils. Estimates of domain status are usually expressed in terms of a percent or fraction of the domain a student knows. Estimates of domain status are of concern when instructional decisions depend on absolute achievement rather than relative achievement. A decision about an individual student's mastery of an instructional objective, for example, is often based on an estimate of that student's domain status: A student is declared to have achieved sufficient mastery if the student scores high enough on a test measuring that objective.

Keeping in mind the distinctions between absolute and relative achievement, and between partial and complete ordering of students, the following types of information about individual test items seem important for classroom test development. The reader should note that, as with other types of information in item analysis, the interpretation of statistical indices of this information will require programming into the microcomputer certain rules of thumb to assist in the decision-making process.

1. The extent to which test items discriminate among students. Regardless of whether a teacher uses a test to measure relative or absolute achievement, the items on the test should contribute information to the total test score in the same algebraic direction. That is, as a group, the higher scoring pupils on the test should have a rather high probability of answering correctly each test item. (This is not to say that each pupil in the higher scoring group will definitely answer correctly every test item, only that there is a propensity to do so.) When a larger proportion of higher scoring students than lower scoring students answer an item incorrectly, a teacher's interpretation of the total test score becomes confused: These negatively discriminating items tell a teacher that the more a student knew (as reflected by the test score) the less are the chances of answering the items correctly. Negatively discriminating items should be examined by a teacher and either revised or not put on the same test with the positively discriminating items.

A decision about which of the positively discriminating items to place on a test depends on (a) the type of achievement being measured (absolute or relative), (b) the nature of the test specifications, (c) the type of decision to be made, and (d) other properties of the items, such as their difficulty levels, and (e) the type of statistical index used to summarize discrimination. These factors are considered in subsequent sections.

2. Difficulty of the item for the class. As we have described previously, item difficulty plays a role in both improving the effectiveness of instruction and in revising a test item. Item difficulty also plays a role in assembling a test since the difficulty levels of the individual

items comprising a test set the difficulty level of the total test. Item difficulty level also sets limits on item discrimination and on the total test reliability. When a test score will be used for partial or complete ordering, item difficulty plays a role in helping to establish the ability level at which a test is most reliable.

3. Relation of an item to the test blueprint and/or the domain. This information is a judgment of the item-domain congruence and/or an indication of where an item fits in the test blueprint (plan). The item-domain congruence judgments have been described earlier in this report. The second type of information is important to the assembly of a classroom test in that it assures the items on the test have sufficient content scope and behavioral breadth for the total test to be content valid.

4. Projection of statistical properties of the total test from the properties of the individual items. If a teacher is assembling a test by selecting items from a pool of previously used items that have known statistical properties, it would be helpful for the teacher to know what to expect in the way the total test will perform. At the minimum, it would be helpful to obtain an estimate of the mean of the test. Other information may be an estimates of the test reliability and standard deviation. This total test information can be estimated from the statistics available on each item. If the items to be used come from an item bank that has been calibrated using a latent trait model, then other total test properties can be described such as the part of the ability continuum on which the assembled test provides the most information.

Review of Statistical Indices Having Potential for Providing the Information Needed for Domain-Referenced Classroom Tests

Having set out in the previous section the information requirements of domain-referenced classroom item analysis, we turn our attention to specific statistical indices which could provide these kinds of information. In this section, we will return to each area previously described, but limit the discussion primarily to various statistical indices.

Improving and Guiding Instruction

In this section we review a number of item statistics that have potential for providing the classroom teacher with the specific kinds of information listed in Table 1 for improving and guiding instruction. By and large, the statistics we review here are considered without regard to their sampling errors. Sampling errors are important to consider in selecting statistics when inferences are made about estimating population parameters or when one seeks to understand the stability of a numerical result when a replication is important such as in an experiment or survey. The numerical values of the indices discussed in this section, when used by the teacher, will be based on a specific set of students and, therefore, when recomputed on data from a new group of students, will likely yield a different numerical value. However, a teacher is interested primarily in working with the group of students at hand at any particular time. Thus, sampling fluctuations are less of concern when the statistical information is to be used to change the students in the sample in some way. (Sampling fluctuations are more of a concern, however, when using item statistics to revise or select items for a test.)

Table 2 provides a list and a brief description of several statistical indices having some potential for providing item data that will serve the various information needs of teachers and which may possibly be computed on a microcomputer. Below we will describe each of these statistics in more detail, pointing to the advantages and disadvanatages of providing them as part of a microcomputer item analysis program for classroom teachers.

Insert Table 2 here

1. Statistical summary of the class' performance on a test item. Table 2 lists six statistical indices which are defined as follows:

$$p_i = \frac{\sum_{a=1}^{N} Y_{ai}}{N}, \; Y_{ai} = 0, 1 \qquad [1]$$

$$\bar{Y}_i = \frac{\sum_{a=1}^{N} Y_{ai}}{N}, \; m_i \leq Y_{ai} \leq 1_i \qquad [2]$$

$$p_i = \frac{\bar{Y}_i - m_i}{1_i - m_i} \qquad [3]$$

$$V_i = \frac{\sum_{a=1}^{N} (Y_{ai} - \bar{Y}_i)^2}{N} \qquad [4]$$

$$P(x_j)_i = \bar{Y}_{1i}, \bar{Y}_{2i}, \ldots, \bar{Y}_{ji}, \ldots, \; m_i < Y_{ai} < 1_i \qquad [5a]$$

$$= p_{1i}, p_{2i}, \ldots, p_{ji}, \ldots, \; Y_{ai} = 0, 1 \qquad [5b]$$

Table 2. Statistical item data potentially useful for helping a teacher improve and guide instruction.

Type of information a teacher could use	Possible statistical indices	
1. Summary of the performance of the class on each item.	p_i	The fraction or percent of the entire class passing a dichotomously scored item.
	$\bar{Y}_i$	The mean item score of the entire class for an item that is scored in a graded or continuous way.
	P_i	The mean item score, $\bar{Y}_i$, expressed as a percent of the maximum possible item score.
	V_i	A measure of the variability of the item scores of the entire class for an item scored in a graded or continuous way.
	$P(\bar{x}_j)_i$	A function that displays average item score for each of j levels of the total test score.
	$P(\Theta)_i$	The item characteristic curve for a dichotomously scored item.

Table 2 (cont.)

2. Discrepancy between the performance a teacher expected of the class and the actual performance of the class	$D1_i = p_i - Ep_i$	The difference between the percent of the entire class actually passing a dichotomously scored item and the percent of the class the teacher expected to pass the item.
	$D2_i = \overline{Y}_i - E\overline{Y}_i$	The difference the actual mean item score of the entire class and the mean item score the teacher expected the class to obtain.
	$D3_i = P_i - EP_i$	Similar to the above difference except P_i is the mean item score expressed as a percent of the maximum possible item score.
	$D4_i = p_i - \gamma_i$	The difference between the percent of the entire class actually passing a dichotomous item and the estimated percent passing the same item in a suitable norm group (e.g., percent passing in the district or percent passing in the state).

Table 2. (cont.)

		$D5_i = P(x_j)_i - EP(x_j)_i$	Difference between the actual average item score and expected average item score for each of j levels of the total test score.
		$D6_i = A_i/M_i$	Ratio of actual discrepancy to maximum discrepancy between students' choices among options of a multiple-choice item (Huynh, 1983).
3.	Unusual pattern of responses on a test for a student	C_a	Modified caution index of Harnisch and Linn (1981).
		$_a r_{perbis}$	Personal biserial correlation (Donlon & Fisher, 1968). The biserial correlation between a person's item responses and the difficulties of the corresponding items, assuming a normal distribution underlying the item responses.

Table 2. (cont.)

	$r_{a\ perptbis}$	Personal point-biserial correlation (Brennan, 1980, cited in Harnisch & Linn, 1981). The product moment correlation between the item scores for a person and the corresponding item difficulties.
	NCI_a	Norm conformity index (Tatsuoka & Tatsuoka, 1982), a measure of the degree of consistency between an individual's response pattern and the ordering of the items in a norm group.
	v_a	Person fit statistic for Rasch model (Wright and Stone, 1979).
4. Hierarchical ordering of the items on a test.	IRSA matrix	Item relation structure analysis matrix (Tatsuoka & Tatsuoka, 1981) is used as a basis for ordering the test items in a hierarchical directed graph.

Table 2. (cont.)

5.	Change in a class' performance after instruction.	${}_iD_{postpre}$ ${}_iP_{post} - {}_iP_{pre}$	Difference between the percent of pupils answering the item correctly before and after instruction (Cox & Vargas, 1966).
		${}_iD_{ingain} = {}_iP_{01}$	Percent of students who answered the item incorrectly on pretest but correctly on posttest (Roudabush, 1973).
6.	Summary of the seriousness of pupils' errors on an item	$P(r_{ji})$	Proportion of students committing each seriousness level of error, r_{ji}.
		$\bar{r}_{\cdot i}$	Mean rating of the seriousness of errors for the entire group of students on a particular item.
7.	Summary of the types of errors committed on an item.	$P(t_{ji})$	Proportion of students committing each type of error, t_{ji}.

$$
P(\Theta)_i = \begin{cases}
\dfrac{e^{Da_i(\Theta - b_i)}}{1 + e^{Da_i(\Theta - b_i)}} & [6] \\[2ex]
\dfrac{e^{Da_i(\Theta - b_i)}}{1 + 2^{Da_i\,(\Theta - b_i)}} \quad , Y_{ai} = 0, 1 & [7] \\[2ex]
c_i + (1 - c_i) \left[\dfrac{e^{Da_i(\Theta - b_i)}}{1 + e^{Da_i\,(\Theta - b_i)}} \right] & [8]
\end{cases}
$$

Where in the above formulas:

N = number of students taking the test

Y_{ai} = the score of the <u>ath</u> student on the <u>ith</u> item on the test

m_i = the lowest possible score a teacher could assign on the <u>ith</u> item

l_i = the highest possible score a teacher could assign on the <u>ith</u> item

$\bar{Y}_{ji}$ = the mean score of the <u>jth</u> subgroup of the class of students on the <u>ith</u> item (e.g., the lower third)

p_{ji} = the percent of the <u>jth</u> subgroup of the class of students answering the <u>ith</u> item correctly

a_i, b_i, c_i, D, e = the parameters and constants of the family of latent trait models based on a logistic ogive (see, e.g., Lord, 1980)

We note immediately that if Y_{ai} is a score on an item graded zero or one, then $p_i = \overline{Y}_i = P_i$. When items are scored in a more continuous fashion, p_i is not used and $\overline{Y}_i \neq P_i$.

An advantage of using statistics [1], [2], or [3] is that they provide a single summary number that can capture the performance of a class of students on a particular test item. A disadvantage, of course, is that these statistics do not provide a summary of how different types or groups of students performed: for example, how the lower third of the class performed compared to how the upper third performed. Thus, some information that is possible to obtain from the item is lost.

An advantage of [3] is that it expresses the average performance of the class on a scale whose range is 0.00 to 1.00. This index, which is described in Whitney and Sabers (1970), is interpreted as the percent of the distance from the lowest to the highest possible score that the class' average item score represents. Thus, values of P_i near 1.00 mean that, on the average, students knew most of the material required by the item, whereas values of P_i near 0.00 mean that generally students did poorly on the item. This interpretation is consistent with the interpretation given to p_i when p_i is used with dichotomously scored items. A disadvantage of [3] is that a teacher may lose a sense of the absolute level of the scores. For example, a $P_i = .80$ may mean $\overline{Y}_i = 4.2$, 4.0, 3.4, or 3.2 depending on whether $(m_i, 1_i) = (5, 1)$, $(5, 0)$, $(4, 1)$ or $(4, 0)$, respectively. This confusion can be lessened, perhaps, by making sure that $\overline{Y}_i$ is available to the teacher upon request. The relationship between P_i and $\overline{Y}_i$ is as follows:

$$\overline{Y}_i = P_i(1_i - m_i) + m_i \qquad [3a]$$

The indices V_i, $P(x_j)_i$, and $P(\Theta)_i$ all describe in one way or another how the members of a group differ from each other. The item variance, V_i, has the advantage of being a single number that measures the spread of individuals. It has serious interpretive problems, however, from the teacher's viewpoint. To understand V_i a teacher would have to have a sense of the concept of a variance--a concept most teachers do not have. Second, V_i is not expressed on the same scale as the item score, so the square root of V_i would need to be taken. Third, one usually cannot compare the variance or standard deviation of one test item to that of another test item because of scaling differences.

The functions [5a, b] and [6], [7], and [8] provide the maximum information about how students perform on an item, but do so in noncomparable ways. The latent trait models represented by $P(\Theta)_i$ describe the probability of each student answering the <u>ith</u> item correctly and thus these models provide a profile of the item performance over the full range of ability. But these latent trait item characteristic functions have serious drawbacks when used to describe the performance of a particular group of students. First, they express performance as a function of an arbitrary ability score, Θ, a concept with which teachers are unfamiliar. Second, they cannot, and probably should not, be calculated on sample samples of data such as are available to the teacher for the class at hand. Third, if a teacher uses items from an item bank (or other source) which are already calibrated using one of the latent trait models, the display of an item characteristic curve can be easily misinterpreted. The ability distribution (on the Θ-scale) of the particular students in the class is unknown and, hence, the teacher has no way of knowing to which parts of the ability scale to refer in order to interpret the item.

Finally, teacher's tests are unlikely to be long enough or homogeneous enough to routinely use latent trait procedures to estimate students' abilities on the Θ-scale.

The function $P(x_j)_i$ may serve a purpose in helping the teacher understand how different levels of students performed on the test. Function [5a, b] expresses the average item performance in terms of the total test score. (Sometimes this is called the item-test regression curve (e.g., Lord, 1980).) However, a teacher could use a scale other than the total test score in this function: It might be reasonable, for example, to use pupils' grades (A, B, C, etc.) in the subject from the previous marking period.

Since experience indicates that the function [5a, b] will not be regular for small samples when they are based on each possible value of the total score, it is likely that a useful form of [5a, b] is to group the students in some way and then show the average item performance for each of these subgroups. Upper half versus lower half is likely to be a too coarse and an uninformative interval width. We recommend dividing the class of students into either thirds (lower, middle, and upper scoring students) or fourths (using quartiles as the dividing points) of the class, if the number of students is between 25 and 40. Larger classes could be sectioned into fifths (using quintiles).

<u>Summary</u>. Table 3 summarizes our recommendations based on the rational consideration put forth above. These recommendations are further reviewed and modified as a result of some empirical (Monte Carlo) studies reported later in this report.

Insert Table 3 here

able 3. Recommended item statistics for helping a teacher improve and guide instruction: Summarizing the performance of the class on each item.

	Basic: Should be included in every item analysis program, if at all possible.		Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
pe of item oring:	Routinely present to or interpret for a teacher on every test item.	Make available to teachers upon their request only.		
chotomous ($_{ai} = 0,1$)	P_i $P(x_j)_i$			V_i $P(\theta)_i$
aded or ontinuous ($_i \leq Y_{ai} \leq 1_i$)	P_i $P(x_j)_i$	$\overline{Y}_i$		V_i

e: See the text for definitions, formulas, and explanations.

Note. We should mention here that we are not recommending that b_i, the difficulty parameter of a latent trait model, be reported for purposes of improving and guiding instruction. The considerations which led us not to recommend $P(\Theta)_i$ have led us not to recommend reporting b_i for purposes of guiding and improving instruction.

2. Discrepancy between the performance a teacher expected of the class and the actual performance of the class. The statistics listed in Table 2 are defined as follows:

$$D1_i = p_i - Ep_i,\ Y_{ai} = 0,\ 1 \qquad [9]$$

$$D2_i = \overline{Y}_i - EY_i,\ m_i \leq Y_{ai} \leq 1_i \qquad [10]$$

$$D3_i = \bar{P}_i - E\bar{P}_i,\ m_i \leq Y_{ai} \leq 1_i \qquad [11]$$

$$D4_i = p_i - \tilde{\phi}_i,\ Y_{ai} = 0,\ 1 \qquad [12]$$

$$D5_{ji} = P(x_j)_i - EP(x_j)_i \qquad [13]$$

$$D6_i = A_i/M_i \quad ;\ Y_{ai} = 0,\ 1 \qquad [14]$$

In the above formulas p_i, Y_{ai}, $\overline{Y}_i$, m_i, 1_i, $\bar{P}_i$, and $P(x_j)_i$ have been defined previously in Formulas [1]-[5a, b]. We note the following clarifying definitions:

E = expectation or "expected value of", but this is not necessarily a mathematical expection (see below).

$\tilde{\phi}_i$ = an estimate of the proportion passing item i in a norm group (e.g., a school district or children at the same grade level in the state's norms)

$$A_i = \sum_{l=1}^{k_i} \left| t_{li} - s_{li} \right| \qquad [14a]$$

= actual discrepancy

k_i = the number of options in ith multiple-choice item

t_{li} = the number of students the teacher expects to choose Option l of Item i

s_{li} = the actual number of students who chose Option l of Item i

$s_{hi} = \min (s_{li})$

$$M_i = N_i - s_{hi} + \sum_{l \neq h} s_{li} \qquad [14b]$$

= maximum possible discrepancy

All of indices [9] through [14] require a method for a teacher to use to specify how the students in the class are expected to respond to a particular test item. There are two general ways for a teacher to arrive at this expected performance for the class at hand: (a) use subjective judgment based on past experience with these students, and (b) use empirical information and a statistical estimate. Although not dismissing statistical estimates as inappropriate for the purposes at hand, we are inclined to favor the judgmental approach for most instructional purposes, especially for Ep_i, $\overline{EY}_i$, and EP_i in Equations [9], [10], and [11]. We would like teachers to become directly involved in "messing around" with data from their students. We feel it serves important instructional purposes for a teacher to compare his or her expectations of pupils on particular test items measuring instructional objectives the teacher operationalizes via test items. If a disparity exists between a teacher's expectations and the pupils' performance, we believe this will be a powerful motivator for the teacher to explore further for an explanation.

We could, of course, use various statistical procedures (regression, Bayesian analysis, etc.) to estimate how a teacher's class will perform. Such estimates, will almost certainly contain errors of estimation due to scedasticity. Further, these estimates would be created by a microcomputer program in a "black box" atmosphere about which a teacher is likely to understand very litte. It may well be that statistical estimates are more efficient, sufficient, and consistent, but their impact on a teacher's behavior is likely to be less in such black box situations than if the teacher was more personally involved.

Equations [9], [10], and [11] correspond to Equations [1], [2], and [3], respectively. We have already indicated the advantages and disadvantages of p_i, $\overline{Y}_i$, and P_i, and have indicated our recommendations with respect to each (see Table 3). To use [9]-[11] in an item analysis program, a teacher would be asked to specify at the time the test is assembled (before it is administered) the anticipated class performance on each item.*

To be consistent with our previous recommendations, we would recommend using [9] and [11] whenever [1] and [3] are used. We anticipate, however, that [11] will be difficult for teachers to use because it requires a two-step process: first estimate $\overline{Y}_i$ and then estimate the percent of $(l_i - m_i)$ which $\overline{Y}_i$ represents. To avoid this complication, we suggest that in an interactive microcomputer program, the teacher be asked to specify m_i, l_i, and $E\overline{Y}_i$;

---

*If experience indicates that this is too tedious to do for each item, various alternatives could be used. For example, the teacher could be asked to specify a single value $Ep_{.}$, $E\overline{Y}_{.}$, or $EP_{.}$ that would represent the items and the deviation of each item from this single value could be computed. Another alternative is to write the program so that EP_i, $E\overline{Y}_i$, or EP_i can be specified for only a few (say the most important) items, rather than all of the items.

then the computer can compute EP_i via:

$$EP_i = \frac{E\overline{Y}_i - m_i}{l_i - m_i} \qquad [11a]$$

It sometimes occurs that a teacher will use test items that have to be administered to a broader group of students of which the students in the teacher's class are a subgroup. For example, a school district may have developed a series of mastery tests and have item analysis data available; a state may use a state-wide assessment program for which test items and data are released to the teacher; or a teacher may be using an item bank that contains items calibrated by one of the latent trait models. In cases such as these, it would be instructive to the teacher to compare the performance of the students in the teacher's class to the performance of similar students in the broader group. Equation [12] specifies this comparison for items score dichotomously.

It is unlikely that a teacher would have access to items scored in a more continuous way since most large scale testing programs use multiple-choice items. An exception to this practice is the situation in which writing samples are taken and graded, a more frequent practice among school districts in recent years. (It might be noted that in many countries outside of the United States, essay tests are more frequently used than multiple-choice tests.) Although we do not treat the case of nondichotomously scored items here, we note that [12] could be adapted easily to accomodate essay tests.

The quantity $\tilde{\phi}_i$ can be obtained in several ways. Computer printouts and technical reports obtained through the testing office of a school district would normally contain the proportion of students in the broader group passing each test item. These can be entered into the microcomputer. If the items a teacher uses measure a unidimensional latent trait and have been calibrated via a latent trait model, then a more refined technique could be used to obtain $\tilde{\phi}_i$. This is explained below.

A teacher's class will vary from year to year in average ability (as expressed on the latent trait scale Θ). If a teacher compares p_i for his or her class with the corresponding index for the broader group (district, state, etc.), the comparison may be somewhat misleading in that the more appropriate reference group would be "students with the same ability as those in this class" rather than "students in general". In effect, the teacher would like to hold ability constant and compare this class to those of similar ability. This can be done via the test characteristic curve and item characteristic curve of latent trait theory in a way that keeps the resultant information in a metric the teacher can understand. The procedure is as follows:

(1) Determe the test characteristic curve for the test.

(2) Compute the mean raw score, $\overline{X}_k$, on the test for the teacher's class, k.

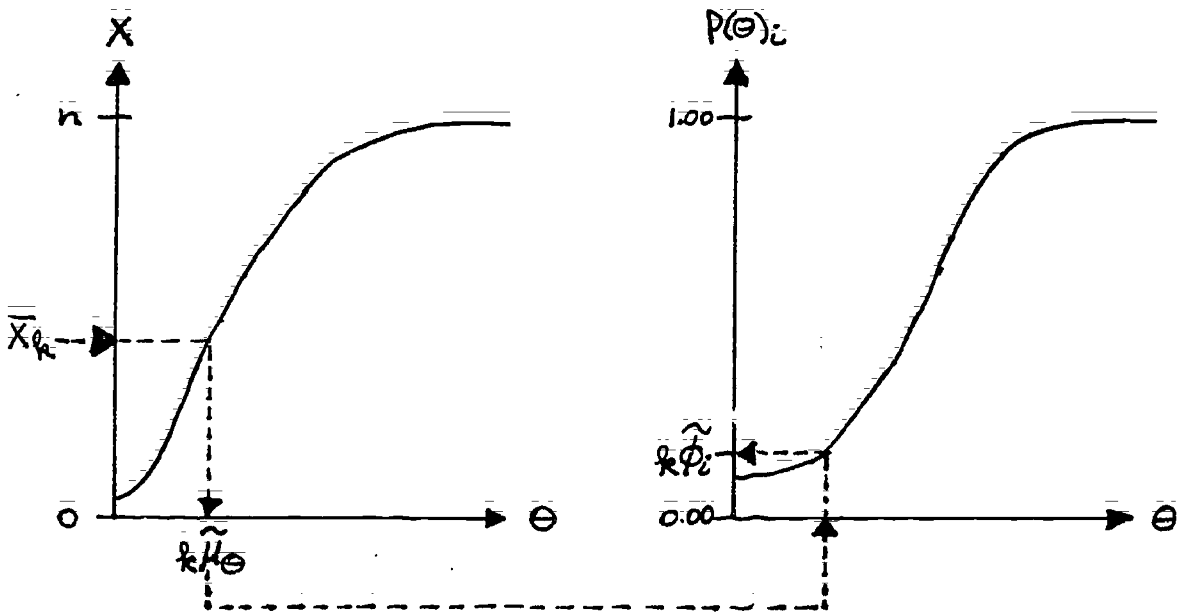
(3) Use this raw score mean and the test characteristic curve to estimate the mean ability level, ${}_k\tilde{\mu}_\Theta$, of the students in this teacher's class.

(4) Use the estimated mean ability level of the class with the item characteristic curve, $P(\Theta)_i$, to estimate the quantity ${}_k\tilde{\phi}_i$ for this class. This is the proportion correct for item i in the norm group for those with ability equal to $\tilde{\mu}_\Theta$.

The above procedure is illustrated in Figure 1.

Insert Figure 1 here

Equation [13] describes the discrepancy between the expected performance of different levels of students with their actual performance on the ith item.



A. Test characteristic curve

B. Item characteristic curve

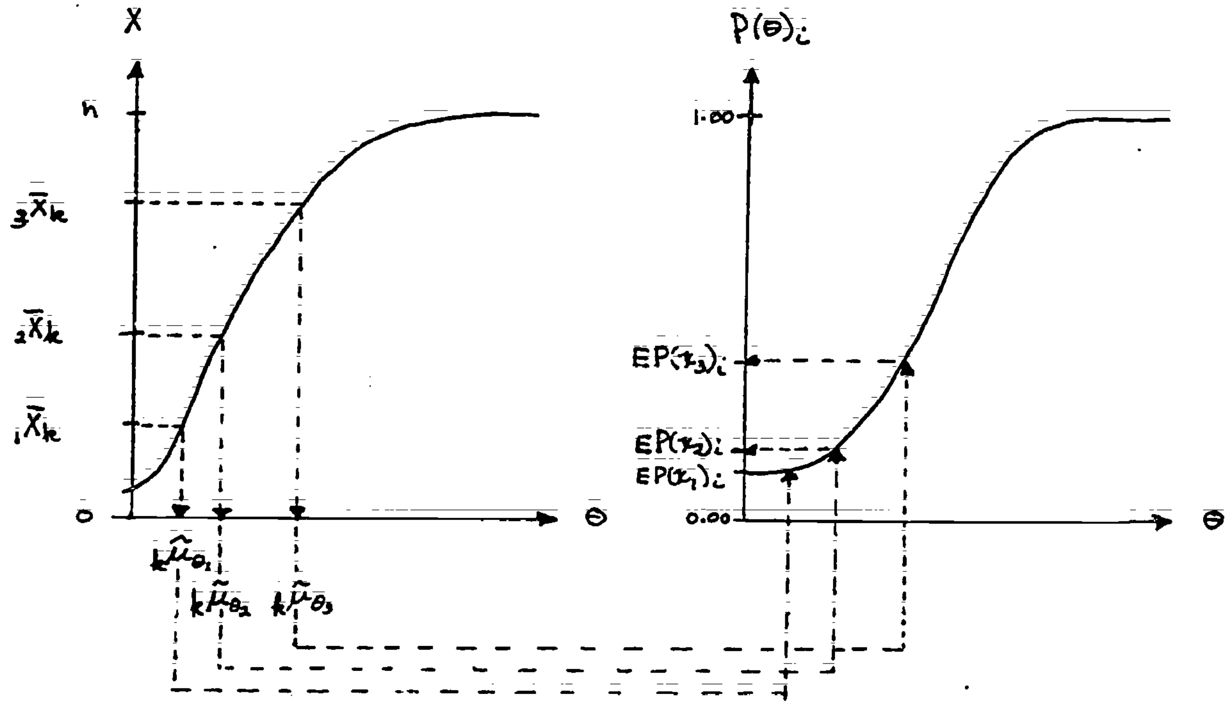
Figure 1. Illustration of the procedure used to estimate the proportion of the norm group (with the same ability as the teacher's class) answering correctly the <u>ith</u> item.

This equation corresponds to [5a, b]. As was indicated when $P(x_j)_i$ was discussed, it seems appropriate to divide the class into thirds or fourths (thus j = 1, 2, 3 or j = 1, 2, 3, 4) unless the group is very large ($\geq$ 50). To use [13] a teacher would be required to specify the expected average item score (either p_{ji} or $\overline{Y}_{ji}$) for each of the j levels of students. If precalibrated latent trait items are used, then $P(x_j)_i$ could be obtained for a particular norm group using the test characteristic and item characteristic curves in a manner similar to that described for Equation [12] and shown in Figure 1. Figure 2 illustrates this procedure for [13] when the class is divided into thirds.

Insert Figure 2 here

Huynh (1983) recently suggested another index of item discrepancy which for multiple-choice items is defined by Equation [14]. This index is a ratio of the actual discrepancy between students' performance and a teacher's expectations to the maximum possible discrepancy for a particular set of teacher's expectations. This index requires the teacher to specify for each option l, of multiple-choice Item i, the number of students expected to choose that option. The statistic represented by [14] close to 0.00 represent agreement between the pattern of student responses to a multiple-choice item and the teacher's expectations; values close to 1.00 represent disagreement.

An advantage of [14] is that it permits teachers to specify a pattern of responses to multiple-choice items. Thus, teacher's would have to consider the nature of each option in relation to the students at hand. If the options were based on specific kinds of errors or misconceptions, the teacher would need to consider the number of students in the class likely to make each error type. While such fine-grained considerations as the expected number of students who would commit each type of error would seem to be a powerful means



A. Test characteristic curve B. Item characteristic curve

Figure 2. Illustration of the procedure used to estimate the expected proportion of the norm group (at each level of ability as the teacher's class) answering correctly the ith item.

of improving a teacher's awareness of student performance and be helpful in guiding instruction, we list several disadvantages: (1) many teachers do not use multiple-choice items, (2) teacher-made multiple-choice items may not be based on particular error types, (3) teacher's may not have the patience to carefully consider for each item the number of students likely to choose each option, and (4) teachers may question the usefulness of such detailed specififcations for every item. These practical, human engineering considerations lead us not to recommend the computation of [14] for purposes of item analysis programs designed to improve and guide instruction. However, [14] does seem to be a useful index for measuring the extent to which pupil responses deviate from a particular pattern. For example, an adaptation of [14] may be useful for detecting guessing patterns. (See a subsequent section of this report.)

Summary. Table 4 summarizes our recommendations based on the rational considerations described above. These recommendations are further reviewed and modified as a result of the Monte Carlo studies reported later in the report.

Insert Table 4 here

3. Unusual performance of a student on a test. The statistics listed in Table 2 are defined as follows:

$$C_a = \frac{\sum_{i=1}^{n_{a.}} (1 - Y_{ai})\, n._i - \sum_{i=n_{a.}+1}^{I} (Y_{ai} n._i)}{\sum_{i=1}^{n_{a.}} n._i - \sum_{i=I+1-n_{a.}}^{I} n._i}, \quad Y_{ai} = 0, 1 \qquad [15]$$

ble 4. Recommended item statistics for helping a teacher improve and guide instruction: Identifying discrepancies between the performance a teacher expected and the actual performance of the class.

e of item ring:	Basic: Should be included in every item analysis program, if at all possible. Routinely present to or interpret for a teacher on every test item.	Make available to teachers upon their request only.	Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
hotomous $_i = 0,1)$	$D1_i = p_i - Ep_i$ $D5_i = P(x_j)_i - EP(x_j)_i$		$D4_i = p_i - \tilde{\phi}$	$D6_i = A_i/M_i$
ded or ntinuous $\leq Y_{ai} \leq 1_i)$	$D3_i = P_i - EP_i$ $D5_i = P(x_j)_i - EP(x_j)_i$	$D2_i = \bar{Y}_i - E\bar{Y}_i$		

: See the text for definitions, formulas, and explanations.

$$ {}_a r_{perbis} = \left[\frac{{}_a\bar{\Delta}_R - {}_a\bar{\Delta}_1}{S_{{}_a\Delta_R}} \right] \cdot \left[\frac{n_{a\cdot}/n_{aR}}{u_a} \right] , \; Y_{ai} = 0, 1 \qquad [16] $$

$$ {}_a r_{perptbis} = \frac{\sum_{i=1}^{I} (Y_{ai} - \bar{Y}_{a.})(p_i - \bar{p}.)}{{}_aS_Y \; S_p} , \; Y_{ai} = 0, 1 \qquad [17] $$

$$ NCI_a = 2S_a/S - 1 \; , \; Y_{ai} = 0, 1 \qquad [18] $$

$$ v_a = \frac{1}{I} \sum_{i=1}^{I} \left[\frac{(Y_{ai} - P(\Theta_a)_i)^2}{P(\Theta_a)_i(1 - P(\Theta_a)_i)} \right] , \; Y_{ai} = 0, 1 \qquad [19] $$

Where in the above formulas:

$i = 1, 2, \ldots, I$ items

$a = 1, 2, \ldots, N$ examinees

$n_{a.} = X_a$ = total score (total correct) for examinee a

$n._i$ = total number of students answering Item i correctly

$Y_{ai} = 0, 1$ = the item score for examinee a on Item i

$\Delta_i = 4z_i + 13$ = normalized difficulty index

z_i = inverse normal transformation of p_i.(p_i = proportion of the class answering Item i correctly)

$$_a\bar{\Delta}_I = \frac{\sum_{i=1}^{I} Y_{ia}\Delta_i}{n_{a.}}$$

= mean Δ for the items Student a marked correctly

$_a\bar{\Delta}_R$ = mean Δ for the items Student a reached

$$= \frac{\sum_{i=1}^{I} \Delta_i}{I} \text{ if Student a attempted all items}$$

$S_{_a\Delta_R}$ = standard deviation of the $_a\Delta_R$

$$= \frac{\sum_{i=1}^{I} (\Delta_i - {_a\bar{\Delta}_R})^2}{I} \text{ if the student attempted all items}$$

n_{aR} = number of items Student a reached

= I, if the student attempted all items

u_a = ordinate of the normal curve which divides the area under the curve into proportions $(n_{a.}/n_{aR})$ and $[1 - (n_{a.}/n_{aR})]$

$\bar{Y}_{a.} = X_a/I$

= mean item score for Person a

$$p_i = \sum_{a=1}^{N} Y_{ai}/N$$

= fraction of the class answering Item i correctly

$$\bar{p}. = \sum_{i=1}^{I} p_i/I$$

= mean item difficulty

$$ {}_aS_Y = \frac{\sum_{i=1}^{I} (Y_{ai} - \bar{Y}_{a.})^2}{I} $$

= standard deviation of Person a's item score

$$ S_{\bar{p}} = \frac{\sum_{i=1}^{I} (p_i - \bar{p}_.)^2}{I} $$

= standard deviation of the item difficulties

S_a = sum of the above-diagonal elements in a dominance matrix for Examinee a when items have been ordered on the basis of p-values from easiest to highest (see Tatsuoka & Tatsuoka, 1982)

S = sum of all the matrix elements in the above mentioned dominance matrix (Tatsuoka & Tatsuoka, 1982)

$P(\theta_a)_i$ = probability of Person a correctly answering Item i as this is predicted from the Rasch model

First, we note that Equations [15]-[19] all apply to dichotomously scored items. Thus, to the extent that classroom tests are not dichotomously scored, these indices will be inappropriate to include in an item analysis program.

Equations [15] and [18], the modified caution index (Harnisch & Linn, 1981) and the norm conformity index (Tatsuoka & Tatsuoka, 1982), respectively, are based on the pattern of an examinee's responses to items when the items have been arranged in order of difficulty from lowest to highest. If examinees respond to the items in a manner consistent with their total test scores, the zero/one elements of an examinee-by-item matrix should appear much as a

Guttman (1950) scalogram. That is, when examinees are arranged in order of total test score and items are arranged in order of difficulty, examinees with high test scores should exhibit an unbroken string of 1s, while examinees with very low scores should have a long unbroken string of 0s. High scoring examinees who break this pattern by responding incorrectly to very easy items (or low scoring examinees who break it by answering difficult items, should be identified via [15] or [18] as performing inconsistently. Pupils so identified by a statistical index can be brought to the attention of a teacher who can seek an explanation.

In recent empirical studies of these two indices Harnisch and Linn (1981) and Rudner (1983) found they correlated quite highly with each other when they were computed on the same data. The modified caution index, however, correlated less with the total score than did the norm conformity index (Harnisch & Linn, 1981). When the purpose of using an index is to identify persons with unusual response patterns, it is undesirable for that index to be confounded (and hence correlated) with the total test score. Using the correlation with the total test score as a criterion, the modified caution index, [15], would be preferred over the norm conformity index for our purposes.

Equations [16] and [17] are correlational indices: the personal biserial (Donlon & Fischer, 1968), [16], and the personal point biserial, [17]. The empirical studies by Harnisch and Linn (1981) and Rudner (1983) demonstrate that these indices are highly correlated with each other when computed on the same data. Harnisch and Linn also found that both indices were correlated with the total score to an unacceptable degree and that sometimes the personal point biserial had a nonlinear relationship to the total score.

Using simulated data, Rudner found that [16] and [17] identified aberrant score patterns of examinees more frequently than [15] and [18] when a 45 item "classroom test" was simulated; [15] and [18] seemed to identify aberrant score patterns more frequently than [16] and [17] when a longer, 80-item, "commercial test" was simulated. Thus, although all four of the indices are intercorrelated they do not identify unusual score patterns with equal effectiveness. In an unpublished study Meyers (reported in Donlon and Fischer, 1968) found that if test items are generally difficult for a group of students, those students who had a better command of the subject (as a result of having taken a course in the subject) tended to have somewhat lower personal biserials.

Finally, Donlon and Fischer point out that the item difficulties (ETS Δs) used in [16] should be derived from a sample independent of the one of which the examinee whose personal biserial correlation is being computed, otherwise the personal biserials will tend to be higher because the person is part of the sample.

Classroom tests are typically short: shorter than the 45 item test Rudner studied. Further, a teacher may not have available item difficulties from previous administrations of the items. Finally, the typical class size, 25-35, is a rather small sample and would surely accentuate chance dependencies in the data. These considerations, along with the finding of researchers such as Harnisch, Linn, and Rudner, lead us to conclude that [16] and [17] should not be used in a classroom item analysis program.

Equation [19] is the unweighted person fit statistic from the Rasch model. It would be used only when the teacher had access to items previously calibrated by this model. This statistic compares a person's actual responses to the items, Y_{ai}, with the person's average (expected) response, $P(\Theta_a)_i$, when the person's ability score, Θ_a, is known. This squared deviation standardized by dividing by the variance of the expected responses for that ability level, summed over all items, and averaged. Rudner (1983) indicates that [19] is more influenced by student responses to very easy and very difficult items. His empirical investigation indicated that [19] was not a very accurate identifier of aberrant response patterns for a simulated classroom test of 45 items, but that [19] did function well with an 80-item simulated commercial test. Given these findings we believe that insufficient evidence exists to include this index in an item analysis package for use with short classroom tests of the type typically encountered in schools. Therefore, we do not recommend including it in a typical item analysis package.

Summary. Table 5 summarizes our recommendations for this section. These recommendations are further reviewed and modified as a result of empirical studies reported in a later section.

Insert Table 5 here

4. Hierarchical ordering of the items on a test. A number of techniques exists for constructing hierarchical orderings among items (e.g., Airasian & Bart, 1973; Bart & Krus, 1973; Wise, 1981; Takeya, 1981). Tatsuoka and Tatsuoka (1981) reviewed these techniques and found Takeya's to be most appealing because it is "... mathematically elegant, and it has algebraic relations with Loevinger's homogeneity [1948] index, Mokken's [1971] index ..., caution index (Sato, 1975), and Cliff's [1977] index C_{t3}" (p. 1).

Takeya's procedure (cited in Tatsuoka & Tatsuoka, 1981) defines an order structure by determining the expected proportions of dominance relationships between two items. This procedure is called item relations

Table 5. Recommended item statistics for helping a teacher improve and guide instruction: Identifying unusual performance of a student on a test.

ype of item coring:	Basic: Should be included in every item analysis program, if at all possible. Routinely present to or interpret for a teacher on every test item.	Basic: Make available to teachers upon their request only.	Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
ichotomous ($Y_{ai} = 0,1$)	C_a			$_a r_{perbis}$ $_a r_{perptbis}$ NCI_a V_a
raded or continuous ($m_i \le Y_{ai} \le 1_i$)				

te: See the text for definitions, formulas, and explanations.

structure analysis (IRSA). "The advantage of using IRSA is (according to Takeya) that it enables us to see a cognitive aspect of a student's performance on the items to a certain extent. Since it generates a digraph representing the hierarchical structure of the items, it will--at the very least--allow us to check the extent to which we have succeeded in constructing problems that require a hierarchically specified set of skills for solving them" (Tatsuoka & Tatsuoka, 1981, pp. 1-2). Tatsuoka and Tatsuoka used the procedure to successfully construct a digraph of the structural relations among a set of 24 items measuring knowledge of addition and subtraction of fractions.

Although the results reported by Tatsuoka and Tatsuoka are encouraging, more experience is needed with microcomputer computation in order to decide on the practicallity of the IRSA approach. Specifically, computer memory requirements and speed of computation need to be determined. Therefore, we recommend the technique be used only if the particular microcomputer to be used is capable of handling the needed computations.

We note that an IRSA matrix for a particular set of items is subject to errors of sampling students. In order to be meaningful, a hierarchical arrangement of items should apply to a defined population of students rather than only to the particular students at hand. Thus, there should be some sample to sample stability of the IRSA matrix. We know very little about the influence of student sampling on the fluctuations of the IRSA matrix. The nature of the stability of the IRSA matrix should be a topic for further study.

Summary. Our recommendation is summarized in Table 6. We will not provide in this report empirical data to further clarify our recommendation.

Insert Table 6 here

5. Change in a class' performance on an item after instruction. A number of item indices based on a pretest-posttest (or a two group) difference have found their way into the literature on criterion-referenced testing. Among these indices are proportion-based indices such as the (a) pretest-posttest difference (Cox & Vargas, 1966), (b) uninstructed-instructed group difference (Klein & Kosecoff, 1976), (c) individual gain (Roudabush, 1973), net gain (Kosecoff & Klein, 1974), (c) maximum possible (Brennan, 1974), (e) B index (Hsu, 1971; Brennan, 1972), and (f) internal sensitivity (Kosecoff & Klein, 1974). There are correlational approaches as well: (a) item-criterion group partial r (Darlington & Bishop, 1966), (b) item-total change scores (Saupe, 1966), and item-criterion group multiple-correlation (Darlington & Bishop, 1966). Most of these indices have been suggested as types of discrimination indices for selecting items for criterion-referenced tests in a manner similar to discrimination indices previously discussed in the literature in connection with norm-referenced tests. An excellent summary and review of these indices has been provided by Berk (1980).

Our purpose in this section is to consider item indices that provide a teacher with useful information about how a class' performance on an item changed as a result of instruction. We note, however, that we do not recommend that the above indices be used for item selection. Most pretest-posttest types of indices are subject to rather large sampling fluctuations when used with small groups. Second, teachers that blindly follow a statistical rule of thumb for culling items on the basis of the value of a statistical index are likely to be deceived: (a) items not showing change may still represent

Table 6. Recommended item statistics for helping a teacher improve and guide instruction: Identifying a hierarchical ordering of the items on a test.

	Basic: Should be included in every item analysis program, if at all possible.		Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
Type of item scoring:	Routinely present to or interpret for a teacher on every test item.	Make available to teachers upon their request only.		
Dichotomous ($Y_{ai} = 0,1$)			IRSA	
Graded or continuous ($m_i \leq Y_{ai} \leq l_i$)				

Note: See the text for definitions, formulas, and explanations.

important behaviors to be monitored, (b) danger exists in culling from the domain those items which a teacher has not taught well, and (c) items not showing favorable pretest to posttest changes may represent erroneous behavior that pupils have acquired as a result of instruction. Also, as Ebel (1972) demonstrated, quirks in items themselves often lie behind pre- to posttest perfomrance anomalies.

Pretest-posttest indices can be useful to the teacher in identifying those items on which pupils in a particular group perform in unexpected ways. A teacher armed with such information may then decide whether the items require revision or whether the fault lies with the instruction rather than with the item.

Among the most useful of these indices for the specific purpose of improving instruction are: Cox and Vargas (1965), Roudabush (1973), and Kosecoff and Klein (1974). In addition, the index proposed by Brennan (1972) and Hsu (1971) has value in examining items when a meaningful passing (mastery) score can be set. The latter requires special interpretive cautions, however, because it cannot be computed when all or none of the students meet the passing score and because the ideal index is zero. The Cox and Vargas index--the difference in the proportions passing from pretest time to posttest time--is a rough gauge of an item's functioning before and after intervening instruction and is likely to be easily understood by teachers. The Roudabush index--the proportion of pupils answering an item correctly at posttest time who also answered it incorrectly at pretest time--more clearly focuses on changes in pupils' performance and it too can be understood by teachers. The Hsu and Brennan index describes how well an item distinguishes between test passers and nonpassers and, so, is not quite a measure of before and after instruction change.

For purposes of giving a teacher information about changes in a class' performance as a result of instructions we would recommend Cox and Vargas (1966) index, Equation [20], and Roudabush (1973, Equation [21], which are defined by the formulas below.

$$ {}_iD_{postpre} = {}_iP_{post} - {}_iP_{pre} \quad , Y_{ai} = 0, 1 \qquad [20] $$

$$ {}_iD_{indgain} = {}_iP_{01} \quad , Y_{ai} = 0, 1 \qquad [21] $$

In the above formulas the notation is the same as that used in equations [1]-[19] except that:

${}_iP_{post}$ = proportion of the class answering Item i correctly on the posttest.

${}_iP_{pre}$ = proportion of the class answering Item i correctly on the pretest

${}_iP_{01}$ = proportion of the class answering Item i correctly on the posttest but incorrectly on the pretest

Indices [20] and [21] are limited to use with dichotomously score (0 or 1) items. However, we can derive comparable version of these formulas for item that are scored in a graded or more continuous fashion:

$$ {}_iD^*_{postpre} = \frac{{}_{post}\overline{Y}_i - {}_{pre}\overline{Y}_i}{l_i - m_i} \quad , m_i \leq Y_{ai} \leq l_i \qquad [22] $$

$$ {}_iD^*_{indgain} = {}_iP^*_{prepost} \quad , m_i \leq Y_{ia} \leq l_i \qquad [23] $$

$$ {}_iD^{**}_{indgain} = {}_iP^{**}_{prepost} \quad , m_i \leq Y_{ia} \leq l_i \qquad [24] $$

In formulas [22] through [24], we use the notation that follows:

$_{post}\bar{Y}_{\cdot i}$ = average score of the class on Item i when it is administered at posttest time

$_{pre}\bar{Y}_{\cdot i}$ = average score of the class on Item i when is is administered at pretest time

l_i = maximum possible score on Item i

m_i = lowest possible score on Item i

$$_{i}p^{*}_{prepost} = \sum_{b=1}^{C}\sum_{d=c}^{D} {}_{i}p_{bd} \qquad [23a]$$

= proportion of the group failing at pretest and passing at posttest time

b = 1, 2, ..., B

= indexes the score categories on Item i at pretest time

d = 1, 2, ..., D

= indexes the score categories on Item i at posttest time

C = the index number of the minimum passing score on Item i; $1 \leq C \leq D$

$_{i}p_{bd}$ = the proportion of examinees taking Item i that scored in the bth category at pretest time and in the dth category at posttest time

$$_{i}p^{**}_{prepost} = \sum_{b<d}\sum {}_{i}p_{bd} \qquad [24a]$$

= proportion of examinees taking Item i who scored higher at posttest time than at pretest time

Equation [22] converts the difference between the mean pretest and posttest scores of the group to a percent of the maximum possible difference. If simply the mean difference is desired then the numerator of [22] can be reported.

We would recommend, however, that the numerator not be reported to teachers for purposes of guiding instruction. We would recommend instead making available to teachers the actual pretest and posttest means.

If a nondichotomously scored item is assigned a passing score, then Equation [23] can be used to examine the shift in the percent of students passing from pretest to posttest time. This index would be comparable in interpretation to [21]. If no cutoff score or passing score is needed then Equation [24] can be used. This equation computes the percent of students in the class who improved their score on item i from pretest to posttest time. Figure 3 shows the data layout for [22], [23], and [24].

Insert Figure 3

Summary. Our recommendations in this section are summarized in Table 7. We do not provide empirical data on these indices.

Insert Table 7 here

6. Summary of the seriousness of the types of errors pupils committed on an item. In order to provide remedial instruction, a teacher needs to know the types of errors and misconceptions a student has. An item analysis program should provide some way of summarizing for each item the seriousness of the errors committed by the students at hand. In this way, a teacher can focus first on those items for which students' errors seem to be most in need of remediation.

Tatsuoka (1981) developed a quantitative index of the seriousness of errors of different types. Her approach is to use an analog of the norm conformity index, [18], in which students' patterns of erroroneous responses to items are compared against an ideal (and correct) set of steps for solving a problem or completing a task. The index requires (a) specifying a task "tree" of procedural steps (similar to task performance networks developed by cognitive psychologists such as Gagné (1968), Gregg (1976),

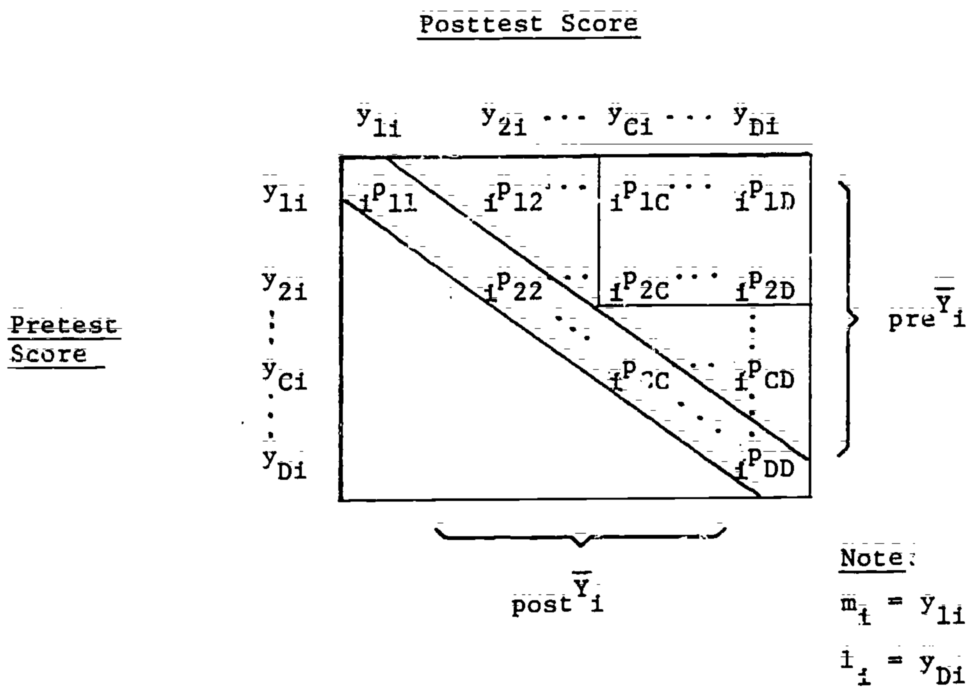


Figure 3. Data layout for Equations [22], [23], and [24]. Equation [22] uses the pre- and posttest means. Equation [23] is the sum of the proportions in the upper right quadrant. Equation [24] is the sum of the elements in the upper triangular portion of the matrix.

Table 7. Recommended item statistics for helping a teacher improve and guide instruction: Identifying changes in a class' performance on an item after instruction

Type of item scoring:	Basic: Should be included in every item analysis program, if at all possible. Routinely present to or interpret for a teacher on every test item.	Basic: Make available to teachers upon their request only.	Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
Dichotomous ($Y_{ai} = 0,1$)	${}_iD_{postpre}$ ${}_iD_{indgain}$			
Graded or continuous ($m_i \le Y_{ai} \le l_i$)	${}_iD^*_{postpre}$ ${}_iD^{**}_{indgain}$	$(\bar{Y}_{post\,i}, \bar{Y}_{pre\,i})$ ${}_iD^*_{indgain}$		

Note: See the text for definitions, formulas, and explanations.

Greeno (1976), and Resnick (1976).), (b) classify pupils' errors with respect to types, and (c) analyzing each error type according to the particular procedural steps that were violated. The Tatsuoka approach is a useful one, as demonstrated by her research, but we believe it is too far ahead of the current capabilities of the typical classroom teacher to be able to develop the procedural steps and analyze them in the way necessary to use her approach. We can conceive of a computer program to do some of this analysis for the teacher once a procedural task network is specified and pupils' responses are entered into the computer. We believe that this would be beyond the practical capability of an item analysis program designed to serve the daily needs of teachers. We would encourage research efforts along the lines of Tatsuoka (1981), however.

What seems more in the realm of possibility is to ask a teacher to rate the seriousness of pupil errors committed on each item and then to summarize these for each item by displaying the frequency with which each degree of seriousness occurs in the class and computing an average of these degrees of seriousness. A teacher would be required to rate the degree of seriousness of the errors committed by each pupil on each item. The indices to be computed are:

$$P(r_{ji}) = (p_{1i}, p_{2i}, \ldots, p_{ji}, \ldots, p_{Ji})_i \qquad [25]$$

$$r_{\cdot i} = \sum_{j=1}^{J} p_{ji} r_{ji} \qquad [26]$$

where

r_{ji} = a teacher's rating of the seriousness of a pupil's error(s) on Item i

$j = 1, 2, \ldots, J$

= indexes the different ratings of a teacher on Item i

p_{ji} = the percent of the class who received an error rating of r_{ji} on Item i

The ratings, r_{ji}, can be assigned by the teacher for nonmultiple-choice items or by the microcomputer if a teacher specifies the seriousness, r_{ji}, for each option of each multiple-choice items.

Insert Table 8 here

7. <u>Summary of the types of errors committed on an item</u>. Instead of, or in addition to, rating the seriousness of each error type, a teacher could classify the errors to each item according to type, t_{ji}. The item analysis program can summarize the percent of the class committing each type of errors. Thus,

$$P(t_{ji}) = (p(t_{1i}), p(t_{2i}), \ldots, p(t_{ji}), \ldots, p(t_{Ji})) \qquad [27]$$

where

$j = 1, 2, \ldots, J$

= indexes the different types of errors

t_{ji} = the <u>jth</u> type of error on the <u>ith</u> item

$p(t_{ji})$ = the percent of the class committing the <u>jth</u> type of error on Item i.

Since t_{ji} is likely to be nonmetric, the mean or average error type has no meaning.

Insert Table 9 here

Table 8. Recommended item statistics for helping a teacher improve and guide instruction: Summarizing the seriousness of the types of errors pupils committed on an item.

Type of item scoring:	Basic: Should be included in every item analysis program, if at all possible.		Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
	Routinely present to or interpret for a teacher on every test item.	Make available to teachers upon their request only.		
Dichotomous ($Y_{ai} = 0,1$)	$P(r_{ji})$ $\bar{r}_{\cdot i}$			
Graded or continuous ($m_i \leq Y_{ai} \leq l_i$)	$P(r_{ji})$ $\bar{r}_{\cdot i}$			

Note: See the text for definitions, formulas, and explanations.

Table 9. Recommended item statistics for helping a teacher improve and guide instruction: Summarizing the types of errors committed by students on an item.

Type of item scoring:	Basic: Should be included in every item analysis program, if at all possible. Routinely present to or interpret for a teacher on every test item.	Basic: Make available to teachers upon their request only.	Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
Dichotomous ($Y_{ai} = 0,1$)	$P(t_{ji})$			
Graded or continuous ($m_i \leq Y_{ai} \leq l_i$)	$P_{(t_{ji})}$			

Note: See the text for definitions, formulas, and explanations.

58

70

Rewriting Individual Test Items

In this section we review a number of item statistics for conveying the information in Table 1 in the category of rewriting or revising a particular test item. We take into account a little more of the sampling distributions of these statistics because if a teacher is revising an item, the teacher expects the item to perform in a certain way in the future. We consider sampling distributions in a more empirical way in a later section of this report.

Table 10 provides a list and a brief description of several statistical indices which have some potential value for providing information for teachers seeking to use pupil data to revise items and which have some possibility of being computed via a microcomputer. Below we describe each of these statistics in more detail, pointing to the advantages and disadvantages of providing them as part of a microcomputer item analysis program for classroom teachers.

Insert Table 10 here

1. Extent of item-objective congruence. If an item does not fit a teacher's instructional objective it should be revised. Item-objective congruence can be judged by a teacher or by a group of teachers. If an individual teacher rates the item-objective congruence we designate it:

$$R_{kji} = \text{the rating of the degree of correspondence of Item } i \text{ to Objective } j \text{ by Teacher } k \qquad [28]$$

We suggest that a rating scale be developed for a teacher to use in which the numbers on the rating scale have verbal anchors describing various degrees of correspondence. An alternative procedure is to use some adaptation of the Mager (1973) scheme for judging item-objective congruence, perhaps quantifying the rating in a manner similar to the error seriousness measure of Tatsuoka (1981). We suggest that this latter approach be further explored, but recommend for the moment that [28] be used as a simple rating as described

Table 10. Statistical item data potentially useful for helping a teacher rewrite individual test items.

Type of information a teacher could use	Possible statistical indices	
1. Extent of item-objective congruence	R_{kji}	An individual teacher's rating of the degree to which an item matches a specified instructional objective.
	I_{ji}	Index of item-objective congruence (Rovinelli & Hambleton, 1977). The average rating of several judges as to whether item i matches objective k.
2. Extent of item-instructional event congruence	R^*_{ki}	An individual teacher's rating of the degree to which an item corresponds to what the teacher taught in class or what the students were expected to study.
	$Mdn_{R^*_{ksi}}$	The median rating of students as to whether the teacher or learning materials taught the content on which the item was based.

Table 10. (cont.)

3. Vocabulary level of an item.	r'_i	Readability grade level of an item as determined from a readability formula.
	r^*_{gi}	Percent of words on a defined grade-level list.
	$\overline{r^{*\prime}_{gi}}$	Mean percent of students at a particular grade-level passing a word-meaning test for the target words in the item.
4. Item difficulty level	p_i	The fraction or percent of the entire class passing a dichotomously scored item.
	$\overline{X}_i$	The mean item score of the entire class for an item that is scored in a graded or continuous way.
	P_i	The mean item score, $\overline{X}_i$, expressed as a percent of the maximum possible item score.
	b_i	The difficulty parameter of an item calibrated via a latent trait model.

Table 10. (cont.)

5.	Item discrimination level	D_i	The net D discrimination index (Johnson, 1951). Difference between the percent passing in the upper and lower scoring pupils in the class.
		$_i r_{bis}$	The biserial correlation between item score and total score.
		a_i	The discrimination parameter of an item calibrated via a latent trait model.
6.	Identification of poor distractors	$D(p_{Uji} - p_{Lji}$	For each option j of item i, the difference between the proportion of upper and lower scoring pupils choosing that option.
		$p_{Lji} = 0$	The option j for which the fraction of lower scoring pupils choosing that option equals zero.
7.	Identification of ambiguous alternatives	$p_{Uji} = p_{Uj'i} > p_{Uki}$	Two options, j and j', for which the same number or percent of upper scoring pupils choose these options and for which these percents are larger than the percents for other alternatives, p_{Uki}.

Table 10. (cont.)

8. Identification of miskeyed items	$\max(p_{Uji}) > p_{Uki}$	The option j may be miskeyed if the percent of the upper scoring group choosing it, $\max(p_{Uji})$ is greater than the percent of the upper scoring group choosing the keyed option, p_{Uki}.
9. Identification of patterns of guessing among knowledgeable students	SSQ_i	Frequency chi-square to testing the goodness of fit of the observed proportion of the upper group choosing each option, p_{Uji}, to a uniform distribution.

above.

Reliability of rating item-objective congruence is gained by having several teachers (or other content experts) judge each item. Hambleton (1980) reviews several methods: (a) rating all possible item-objective pairings (Rovinelli and Hambleton, 1977), (b) rating scale, and (c) matching task. The latter consists of having each teacher attempt to match up the test items on one list with the objectives on another. Items for which there exists a lot of disagreement in matching among the teachers are revised. The rating scale method consists of presenting teachers with a list of test items already matched to objectives and asking teachers to judge the degree of correspondence between each item and its corresponding objective. Items for which the median rating is low and/or for which the variance of the ratings is large are revised. We prefer the rating procedure to the matching procedure because it seems to be a simpler task for teachers and rather straight forward and it asks teachers to judge the extent to which they believe that items already sorted into categories by objectives have been properly sorted. Disadvantages of the rating technique are (a) that someone has to do an initial matching of the items and the objectives and (b) it does not allow every item to be compared to every objective. (It sometimes happens in practice that items will correspond better to objectives for which they were not supposed to match. The rating procedure does not allow for this anomally to be detected.) When the rating procedure is used we recommend that the median rating be the summary index.

$Mdn_{R_{ji}}$ = median rating of the correspondence Item i to Objective j by several teachers [29]

The Rovinelli and Hambleton (1977) procedure for judging item-domain congruence seems to be the most thorough of the three procedures. It requires that each teacher in the group judge each item against each objective and rate each pairing as: +1 if the item definitely measures the objective, 0 if the teacher is undecided about the match, or -1 if the item definitely does not measure the objective. A large number of comparisons are required. If, for example, there are 10 objectives with 3 items per objective, there are 300 (= (10 x 3) items x 10 objectives) pairs to judge. A disadvantage of this technique is that because of the large number of comparisons to be made, it is very time consuming. We prefer it, however, to the rating method if time permits its use because it does allow all items and objectives to be reviewed.

The method is implemented by collecting the ratings and entering them into the formula below. The numerical value of the index obtained from the formula for each test item does not depend on the number of objectives or the number of teachers doing the rating. The index ranges in value from -1.00 to +1.00. A value of 0.00 indicates that teachers cannot agree that Item i matches Objective j; a value of +1.00 indicates that all teachers agree that Item i matches Objective j; and -1.00 indicates that all teachers agree that Item i does not match Objective j. The formula given by Rovinelli and Hambleton is:

$$I_{ji} = \frac{(J - 1) \sum_{k=1}^{K} R_{kji} - \sum_{j=1}^{J} \sum_{k=1}^{K} R_{kji} + \sum_{k=1}^{K} R_{kji}}{2(J - 1)K} \qquad [30]$$

where

i = index number of the item

j = 1, 2, ..., J indexes the objectives

k = 1, 2, ..., K indexes the teachers

R_{kji} = -1, 0, +1

= the rating of the kth teacher of the degree of correspondence of the ith item to the jth objective

When using [29], a cut-off value, C_{ji}, is specified. Any item for which $I_{ji} < C_{ji}$ is revised to match the objective better.

Summary. Table 11 summarizes our recommendations here. We do not provide further empirical data for these indices.

Insert Table 11 here

2. Extent of item-instructional event congruence. An item should be revised if it does not correspond to what the teacher taught or what the students were assigned to study. We call this the item-instructional even congruence. The teacher, the students, or both can be asked to judge the degree to which an item corresponds to the instructional events of the classroom. We list two indices below.

R^*_{kji} = the kth teacher's rating of the degree of correspondence of Item i to what was taught to the students [31]

$Mdn_{R^*_{ksi}}$ = median rating of the students in the kth teacher's class as to whether the material in Item i was taught by the teacher or covered by the materials. [32]

ble 11. Recommended item statistics to use to help a teacher revise individual test items: Extent of item-objective congruence.

	Basic: Should be included in every item analysis program, if at all possible.		Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
e of item ring:	Routinely present to and interpret for a teacher on every test item.	Make available to teachers upon their request only.		
hotomous ($_i = 0,1$)	R_{kji}	$Mdn_{R_{ji}}$	I_{ji}	(see text)
ded or ntinuous ($\le Y_{ai} \le 1_i$)	R_{kji}	$Mdn_{R_{ji}}$	I_{ji}	(see text)

: See the text for definitions, formulas, and explanations.

In order to implement [31] and [32] rating scales need to be developed. We suggest a 4 or 5 point rating scale that has verbal anchors describing various degrees of overlap with instructional events. We also suggest that different scales be developed for students and teachers. Items receiving a low rating by the teacher may need to be revised (e.g., if the item came from a set provided by the textbook publisher), or the teacher may have to alter the instruction. If students do not perceive the item as related to what they were taught or studied (i.e., median rating is low) then a teacher may need to discuss the item with the students before deciding whether to revise it.

Summary. Table 12 summarizes our recommendations.

Insert Table 12 here

3. Vocabulary level of an item. Several indices are suggested in Table 10 to judge the appropriateness of the wording of an item. Several readability formulas exist and some could presumably be implemented via a microcomputer. For each item, one applies a readability formula to obtain the item's readability grade level, r_i. Readability formulas, however, require several long passages to be analyzed (e.g., Fry (1979) or Bormuth (1969)). Even when long passages are analyzed, some reading specialists question the validity of these formulas (e.g., Instructional Objectives Exchange, 1980).

When readability formulas are discounted, about the only alternative left is to use a word list of some type. Word lists attempt to identify the pool of words that are appropriate to use on tests (and other materials) designed for students at a particular grade level. Several approaches have been used to develop word lists (IOX, 1980): (a) tabulating the words

Table 12. Recommended item statistics to use to help a teacher revise individual items: extent of item instructional event congruence.

Type of item scoring:	Basic: Should be included in every item analysis program, if at all possible.		Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
	Routinely present to or interpret for a teacher on every test item.	Make available to teachers upon their request only.		
Dichotomous ($Y_{ai} = 0,1$)	R^*_{ki}	$Mdn_{R^*_{ksi}}$		
Graded or continuous ($m_i \le Y_{ai} \le l_i$)	R^*_{ki}	$Mdn_{R^*_{ksi}}$		

Note: See the text for definitions, formulas, and explanations.

appearing at each grade level in published reading textbooks series (e.g., Taylor, et al., 197?) (b) listing the words at each grade level that students know the meaning of (e.g., Dale and O'Rourke, 1976), (c) tabulating the frequency with which words appear in general reading materials (such as newspapers and magazines) (e.g., Carroll, Davies, and Richman, 1971; Sakiey and Fry, 1979), and (d) some combination of (a) through (c) (e.g., IOX, 1980).

One could tabulate for each item the number of words in the item that are on a particular list at a particular grade-level and convert this to a percent,

$$r^*_{gi} = \frac{n_{gi}}{n_i} \qquad [33]$$

where

n_{gi} = number of words in Item i that are found on the appropriate list of eligible words for that grade level, g.

n_i total number of words in Item i,

If this percent is less than some specified level (perhaps 1.00) the item would be revised.

Some word lists were developed by asking students to check the words they knew the meaning of or by giving students multiple-choice vocabulary tests to determine their knowledge. The percent "passing" each word is then listed. If a test item's words are checked against such a list for a particular grade, and the percent of students in the norm group passing each word recorded then, one index for the vocabulary level of an item would be:

$$\overline{r^*_{gi}} = \frac{\sum_{k=1}^{K_{gi}} p_{kgi}}{K_{gi}} \qquad [34]$$

where

p_{kgi} = percent of the norm group at Grade g knowing the meaning of Word k in Item i

K_{gi} = the number of words in Item i which were located on the particular word list for Grade g

g = 1, 2, ... 12 indexes the grade level

k = 1, 2, ..., K_{gi} indexes the words found in the list

A disadvantage of [33] is that an item may contain words not on the prescribed word list which are either (a) above the grade level intended for the item or (b) suitable for the grade level. Thus, if r^*_{gi} is less th 1.00, no immediate course of action can be recommended except to check the vocabulary. Disadvantages of [34] are that (a) K_{gi} may be quite a bit lower than n_i and (b) the values of p_{kgi} may be based on a norm group that is not appropriate for the local pupils. A disadvantage of both [33] and [34] is that is takes a long time to have a microcomputer check the vocabulary level of an item since each word in the item has to be checked against a long list of suitable or target words. Further, the test item itself would have to entered into the microcomputer (i.e., an item bank would be needed).

Of the two procedures for checking word lists, we recommend [33] since its interpretation is likely to be somewhat easier for teachers. We suggest that items be flagged for teachers and that the words not on the word list be listed (or otherwise identified) for the teacher. Since a computer program for doing this type of word processing and checking may not be feasible for a small microcomputer, we recommend that it not be incorporated into a typical item analysis package destined for a computer with small memory.

<u>Summary</u>. Our recommendations in this area are summarized in Table 13.

Insert Table 13 here

---

4. <u>Item difficulty level</u>. The item difficulty level indices listed in Table 10 are the same ones listed in Table 2 (except there are fewer indices in Table 10). Our recommendations for item difficulty indices are listed in Table 14. These are essentially the same as the recommendations in Table 3. For purposes of identifying items that should be considered for revision, it seems better to look at the overall difficulty of the item in the class.

We would not recommend using b_i, the item difficulty level of a latent trait model, for identifying teacher-made items in need of revision. Latent trait models (Lord, 1980; Rasch, 1960; Wright & Stone, 1979) offer another conception of item difficulty: The point on a number line representing the underlying latent ability at which the slope of the item response curve is maximum. Large samples of pupil responses are needed to calibrate items using latent trait models. While some large school districts have the capacity to calibrate pools of items, most do not. Classroom microcomputers are unlikely to have the kind of computing capacity needed to calibrate items. Further, many classroom teachers would have difficulty because the concepts of latent trait and item response functions are not commonly used. Additionally, there is no compelling educational or psychological reason to believe that single objectives (or other instructional domains) <u>ought</u> to be unidimensional (cf., Nitko, 1974), a requirement needed in order for latent trait theory to be applied. Thus, although items pre-calibrated by latent trait methods can

Table 13. Recommended item statistics to use to help a teacher rewrite or revise items: Vocabulary level of an item

Type of item scoring:	Basic: Should be included in every item analysis program, if at all possible. Routinely present to or interpret for a teacher on every test item.	Basic: Make available to teachers upon their request only.	Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
Dichotomous ($Y_{ai} = 0,1$)			r^*_{gi}	r_i $\overline{r_{gi}}$
Graded or continuous ($m_i \le Y_{ai} \le l_i$)			r^*_{gi}	r_i $\overline{r_{gi}}$

Note: See the text for definitions, formulas, and explanations.

be employed in the classroom, this method is unlikely to be of widespread practical value to teachers in the revision of test items.

Summary. Table 14 summarizes our recommendations. We discuss sampling fluctuations of p_i these statistics in a later part of this paper.

Insert Table 14 here

5. Item discrimination level. The following are the definitions of the statistics listed in Part J of Table 10 along with a few additional ones.

$$D_i = p_{Ui} - p_{Li} \qquad , Y_{ai} = 0, 1 \qquad [35]$$

$$D_i' = \frac{\bar{Y}_{Ui} - \bar{Y}_{Li}}{l_i - m_i} \qquad ; m_i \leq Y_{ai} \leq l_i \qquad [36]$$

$${}_i r_{bis} = \frac{\bar{X}_{1i} - \bar{X}}{S_x} \qquad \frac{p_i(1-p_i)}{y_i} \qquad , Y_{ai} = 0, 1 \qquad [37]$$

$${}_i r_{pbis} = \frac{\bar{X}_{1i} - \bar{X}_{0i}}{S_x} \cdot \qquad p_i(1-p_i) \quad , Y_{ai} = 0, 1 \qquad [38]$$

a_i = discrimination parameter of a latent trait model ; $Y_{ai} = 0, 1$ [39]

In the above formulas:

p_{Ui} = percent of the upper or higher scoring (on the total test) group of students who answer Item i correctly

p_{Li} = percent of the lower scoring (on the total test) group of students who answer Item i correctly

Table 14. Recommended item statistics to use to help a teacher rewrite or revise items: Item difficulty level

Type of item scoring:	Basic: Should be included in every item analysis program, if at all possible.		Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
	Routinely present to or interpret for a teacher on every test item.	Make available to teachers upon their request only.		
Dichotomous ($Y_{ai} = 0,1$)	p_i			b_i
Graded or continuous ($m_i \le Y_{ai} \le 1_i$)	p_i	$\bar{Y}_{.i}$		

Note: See the text for definitions, formulas, and explanations.

$\bar{X}_{1i}$ = mean total test score of the students who answered Item i correctly

$\bar{X}_{0i}$ = mean total test score of the students who answered Item i incorrectly

p_i = difficulty index of Item i defined by Equation [1]

$\bar{Y}_{Ui}$ = mean score on Item i of the upper or higher scoring (on the total test) group of students

$\bar{Y}_{Li}$ = mean score on Item i of the lower scoring (on the total test) group of students

S_X = standard deviation of the total test scores of the students

y_i = ordinate of the normal curve corresponding to the area equal to p_i

Traditionally, item discrimination refers to the extent to which an item is able to differentiate among individuals with various levels of total test performance. Most classroom test construction textbooks recommend using the net D index, [35], for dichotomously scored items because it is easily computed and understood by teachers. Originally proposed by A. Pemberton Johnson (1951), this index has the advantage of describing the fraction of net correct discriminations an item makes (Findley, 1956). Here, a correct discrimination means that the item is answered correctly by a high scoring examinee and incorrectly by a low scoring examinee. (The index has been shown to have good properties when used to select items for purposes of measuring relative achievement (White, Feldt, & Sabers, 1975)).

The use of [35], and any index described in this section for that matter, for the purpose of domain-referenced classroom test item revision is not immediately clear, but as we indicated earlier, if on a particular item low scoring pupils do better than high scoring pupils (i.e., the item discriminates negatively), the item needs to be studied more carefully before the teacher decides to keep it intact or to revise it (e.g., Popham & Husek, 1969). It would be a sensible standard practice to revise items that have discrimination indices below zero regardless of the purpose of the test, unless there is some compelling reason not to.

Equation [36] was suggested by Whitney and Sabers (1970) as a counterpart to [35] for items scored in a graded or continuous manner. It expresses the mean difference between the upper and lower scoring groups as a percent of the distance between the maximum possible score, l_i, and the minimum possible score, m_i, for Item i. An alternate version of [36] which makes its meaning clearer is

$$D_i' = P_{Ui} - P_{Li} \qquad [36a]$$

where P_{Ui} and P_{Li} are computed for the upper and lower scoring groups in a manner similar to Equation [2].

Equations [37] and [38] are correlational indices of item discrimination. For our purposes here, they are considered for the purpose of identifying poorly discriminating item that would be identified and flagged for revision by a teacher. Thorndike (1982) reviews the characteristics of the biserial and point biserial correlations as item discrimination indices (particularly as they relate to selecting items for standardized tests). The point biserial is affected by the item difficulty, p_i, which curtails the possible range of ${}_ir_{pbis}$. Thus its value is confounded with item difficulty. The ${}_ir_{bis}$

is not as confounded with item difficulty, but for small samples and in skewed distributions, its numerical value can go beyond the bounds of +1 and −1. The biserial correlation cannot be used in the standard formulas for estimating total test statistics from item statistics.

The disadvantages of the point biserial and biserial correlations would argue against using them to identify items for a teacher to consider rewriting. Net D, [35], seems to be a more straightforward statistic to compute and interpret to teachers. We note, however, that net D is also confounded with the item difficulty level, p_i (Ebel, 1979).

We would not recommend using the latent trait parameter, a_i, for identifying items for teachers to rewrite, for reasons similar to those offered for not recommending, b_i. It should be noted that if a_i were to be used, its use would be limited to precalibrated items of the two- and three-parameter models, Equations [7] and [8], since in the one-parameter model all a-values are equal.

With the exception of [35] all the above mentioned discrimination indices are used only with dichotomously scored items. Graded or continuously scored items can be analyzed with correlation analogues of the biserial and point biserial correlations, namely, the polyserial and point polyserial correlations (Olsson, Drasgow, & Dorans, 1982), respectively.

Summary. We summarize our recommendations for this section in Table 15. We provide some empirical data on these statistics in a subsequent section.

Insert Table 15 here

6. Identification of poor distractors. The distractions of a multiple-choice item have a specific function: appear as plausible answers to those

Table 15. Recommended item statistics to use to help a teacher rewrite or revise items: Identifying poorly or negatively discriminating items.

Type of item scoring:	Basic: Should be included in every item analysis program, if at all possible.		Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
	Routinely present to or interpret for a teacher on every test item.	Make available to teachers upon their request only.		
Dichotomous ($Y_{ai} = 0,1$)	D_i			$_i r_{bis}$ $_i r_{ptbis}$ a_i
Graded or continuous ($m_i \le Y_{ai} \le l_i$)	D_i'			$_i r_{ptpolyserial}$ $_i r_{polyserial}$

Note: See the text for definitions, formulas, and explanations.

students who do not have the degree of knowledge needed to choose the correct answer to the item. Since it is in the lower group that we would expect to find those lacking the requisite degree of knowledge, we would expect the item data from the lower group to provide information about poorly functioning distractors. One of two definitions of a properly functioning distractor is often used: (a) a distractor is properly functioning if more persons in the lower group than in the upper group choose it and (b) a distractor is properly functioning if at least one person in the lower group chooses it. Improperly functioning distractors are either revised, replaced, or removed from the item. The following equations are consistent with these two definitions:

$$D(d_{ji}) = (d_{1i}, d_{2i}, \ldots, d_{hi}) \quad , Y_{ai} = 0, 1 \qquad [40]$$

$$p_{Lji} = 0 \quad , Y_{ai} = 0, 1 \qquad [41]$$

where

$$d_{ji} = p_{Uji} - p_{Lji} \qquad [41a]$$

$j = 1, 2, \ldots, h$ indexes the options of an h-option multiple-choice item, $j \neq$ correct answer

p_{Uji} = the proportion of the students in the upper scoring group choosing Distractor j of Item i

p_{Lji} = the proportion of the students in the lower scoring group choosing Distractor j of Item i

Equation [40] provides the set of differences between the proportion choosing each distractor. If $d_{ji} \leq 0$, then Distractor j would be flagged for the teacher to consider revising. Equation [41] considers only the lower group and looks for a Distractor j for which $p_{Lji} = 0$. When this criterion is met, the Distractor j is flagged.

Of the two formulas, we prefer [40] since it will identify more distractors for the teacher to review. In particular, d_{ji} may be less than or equal to zero even if some persons in the lower group choose Distractor j. The fact that more upper than lower scoring pupils choose an incorrect option should be brought to the teacher's attention.

Some standardized test developers use the biserial or point biserial correlation between total test score and choosing Distractor j as an index of distractor quality. We do not recommend this for the analysis of teacher-made test items for two reasons: (a) because of those reasons specified previously in connection with the discrimination index, and (b) because when net D is used as a discrimination index, the data are set up in a way to make [40] simple to compute.

We recommend also that p_{Uji} and p_{Lji} be made available to the teacher upon request.

Summary. Table 16 summarizes our recommendations for this section.

Insert Table 16 here

7. Identification of ambiguous alternatives. Here we seek to identify multiple-choice items that contain ambiguous alternatives. Our definition of ambiguous is similar to that of Sax (1980): Two alternatives of a multiple-choice item, j and j′, are said to be ambiguous if the same percent of the upper scoring students choose j and j′; and if this percent is the largest percent among the alternatives. One expression for this relation is:

$$p_{Uji} = p_{Uj'i} > p_{Uki} \ \forall \ (j, k) \text{ of Item } i \quad , Y_{ai} = 0, 1 \qquad [42]$$

Table 16. Recommended item statistics to use to help a teacher rewrite or revise items: Identifying poor distractors

Type of item scoring:	Basic: Should be included in every item analysis program, if at all possible.		Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
	Routinely present to or interpret for a teacher on every test item.	Make available to teachers upon their request only.		
Dichotomous ($Y_{ai} = 0,1$)	D(dji)	p_{Uji} p_{Lji}		$p_{Lji} = 0$ (see text also)
Graded or continuous ($m_i \le Y_{ai} \le l_i$)				

Note: See the text for definitions, formulas, and explanations.

where

$$(p_{U1i}, p_{U2i}, \ldots, p_{Uki}, \ldots, p_{Uhi}) = \text{the proportion of the upper scoring group choosing each distractor}$$

We know of no particular index other than [42] or some function of [42] that is suitable for this purpose.

Insert Table 17 here

8. Identification of miskeyed items. We define miskeying to have occurred when the teacher inadvertantly scores an incorrect alternative as the correct answer to Item i for all students. Under this conditions, the largest percent of upper scoring students would choose the right answer to Item i, but it would be marked wrong. This can be specified as follows:

$$\max(p_{Uji}) > p_{Uki} \quad , \quad Y_{ai} = 0, 1 \qquad [43]$$

where

$j = 1, 2, \ldots, h$ indexes the alternatives to Item i

p_{Uji} = the proportion of the upper group choosing Alternative j of Item i

p_{Uki} = the proportion choosing the keyed alternative, k, of Item i

$j \neq k$

As with Equation [42], we know of no other indices other than, perhaps, simple transformations of [43].

Insert Table 18 here

Table 17. Recommended item statistics to use to help a teacher rewrite or revise items: Identifying ambiguous distractors

Type of item scoring:	Basic: Should be included in every item analysis program, if at all possible. Routinely present to or interpret for a teacher on every test item.	Basic: Make available to teachers upon their request only.	Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
Dichotomous ($Y_{ai} = 0,1$)	$p_{Uji} = p_{Uji} > p_{Uki}$			
Graded or continuous ($m_i \leq Y_{ai} \leq 1_i$)				

Note: See the text for definitions, formulas, and explanations.

103

Table 18. Recommended item statistics to use to help a teacher rewrite or revise items: Identifying ambiguous distractors

Type of item scoring:	Basic: Should be included in every item analysis program, if at all possible.		Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
	Routinely present to or interpret for a teacher on every test item.	Make available to teachers upon their request only.		
Dichotomous ($Y_{ai} = 0,1$)	$\max(p_{Uji}) > p_{Uki}$			
Graded or continuous ($m_i \le Y_{ai} \le l_i$)				

Note: See the text for definitions, formulas, and explanations.

9. <u>Identification of patterns of guessing among knowledgeable students</u>. Here we want to use an index that would allow us to flag an item for a teacher if the pattern of responses to it indicated that students who should know the answer to the item are behaving in a random fashion. Two indices of possible guessing behavior that are consistent with this purpose are the following.

$$RU_i = \frac{\sum -p_{ji} \log_2 (p_{ji})}{-\log_2(1/(h_i-1))} = \frac{U_{observed}}{U_{maximum}} \qquad [44]$$

$$D7_i = \frac{D_{observed}}{D_{maximum}} = \frac{\sum_{j=1}^{h} \left| (n_{U..}/h_i) - n_{Uji} \right|}{n_{U..} - \min(n_{Uji}) + \sum_{j \neq h_i} n_{Uji}} \qquad [45]$$

where

p_{ji} = proportion of the entire class choosing Distractor j of Item i

h_i = the number of alternatives for Item i

n_{Uji} = the number of students in the upper scoring group who chose alternative j on Item i

$n_{U..}$ = number of students in the upper scoring group

Equation [44] is known as the relative uncertainty index and was suggested by Pike and Flaugher (1970). This index takes on values between 0.00 and 1.00 and reflects the extent to which examinees respond in a manner that would produce a flat (uniform) distribution of p_{ji} values over the distractors (wrong answers) of a multiple-choice item. The RU_i index has been used successfully to study guessing patternsn in several standardized tests: the <u>PSAT</u> (Pike &

Flaugher, 1970), the GRE (Pike, 1980), the 3R's (Khampalikit, 1982), and the Joint College Entrance Examination of Taiwan (Hsu & Khampalikit, 1980). It was also used to study a college level classroom test (Hsu & Liou, 1982) but with less success. A major problem that occurs with RU_i is that p_{ji} cannot be equal to zero when computing the log. Thus, if all students can eliminate one distractor, RU_i cannot be computed. This is particularly problematic for classroom tests. A second difficulty with using RU_i as stated in [44] is that is considers all students, not just upper group students. We would expect the lower scoring students to guess on classroom tests and teachers may well encourage them to do so. It is in the upper scoring group of students that we believe we should find patterns indicating that they are responding in a more informed manner. Pike and Flaugher do suggest that [44] be computed for various subgroups, but again as the number of responses to each alternative become fewer, the computation and interpretation of [44] becomes problematic.

Equation [45] is an adaptation of the Huynh (1983) index defined by Equation [14]. Although we found [14] not to be practical for identifying items exhibiting teacher-pupil discrepancies, the adaptation, [45], seems useful for the purposes of this section. We substitute for t_{1i} in [14] the expected frequency of choices for each alternative if the upper group responded randomly ($=n_{U...}/h$). The value of $D7_i$ is near zero if the students do respond randomly and is one if students do not respond randomly. Note that unlike the relative uncertainty index, $D7_i$, considers (a) only the upper scoring group and (b) all alternatives, not just the distractors. We believe these characteristics to be advantages since (a) it is when the upper group begins to guess randomly that a teacher's attention should be drawn to the item and (b) if the upper group is randomly guessing their guesses will include the possibility of choosing the correct alternative as well as the distractors.

Still a third index could be computed (this is given in Table 10), the frequency chi-square for testing the goodness of fit to a uniform distribution of the response pattern of the upper group to all alternatives. This equation is

$$SSQ_i = \frac{\sum_{j=1}^{hi} (n_{Uji} - (n_{Uji}/h_i))^2}{(n_{Uji}/h_i)} \qquad [46]$$

where n_{Uji} and h_i are as defined for [44] and [45]. We believe [46] to be too variable with small n_{Uji} so that if a strictly statistical chi-square criterion is used to decide whether the pattern of responses is uniform the user would be subject to committing a Type II error with high probability.

Summary. Table 19 summarizes our recommendations that Equation [45] be used to identify items where guessing may be occurring among the upper group.

Insert Table 19 here

Selecting Items to Put on a Test

We presented the rationale that will guide our review of item statistics for purposes of improving the total test score properties on pages 13-16. In this section we review several statistical indices and make recommendations concerning which should be included in a microcomputer item analysis program for classroom testing.

We assume in this section that the indices will be used to select items,

Table 19. Recommended item statistics to use to help a teacher rewrite or revise items: Identifying patterns of guessing among knowledgeable students

Type of item scoring:	Basic: Should be included in every item analysis program, if at all possible.		Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
	Routinely present to or interpret for a teacher on every test item.	Make available to teachers upon their request only.		
Dichotomous ($Y_{ai} = 0,1$)	$D7_i$			RU_i SSQ_i
Graded or continuous ($m_i \leq Y_{ai} \leq 1_i$)				

Note: See the text for definitions, formulas, and explanations.

rather than for trying to improve instruction by reviewing items or trying to obtain information about which items need to be revised. We are assuming also that the items have been revised and tried out so that the statistics (or the data for the statistics) are available. Thus, the items exist in some pool or bank and that the item analysis program will compute and interpret certain statistical indices associated with each item.

We assume that a teacher will use different classroom tests for different purposes as we outlined on pages 13 and 14. Among other things this means that item statistics will need to be used in combination in order to select items to put on any particular test. The reader is urged to keep this in mind when reading below, because we initially focus on each category of item statistic separately.

Table 20 lists several item analysis statistics which seem on the surface to be suitable for our purposes here. Below we will review them.

Insert Table 20 here

1. Item discrimination level. The three item discrimination indices in Table 20 have been defined and discussed previously (Equation [35] - [39]) for other purposes. Here we note that it seems most appropriate to use net D as specified in [35] and [36] for most classroom tests in a way that we will describe shortly. We do not recommend the correlational indices ${}_ir_{ptbis}$ and ${}_ir_{bis}$, or their polyserial counterparts.

We do recommend, however, that if the teacher has access to an item bank containing items calibrated on a two-parameter or three-parameter latent trait model (Equations [7] or [8]) that the item discrimination index, a_i, be used. It would not be possible for a small microcomputer to compute a_i for classroom tests, but if a_i were already available, it is possible to create

Table 20. Statistical item data potentially useful for helping a teacher select items to put on a classroom test.

Type of information a teacher could use		Possible statistical indices
1. Item discrimination level	D_i	Same as Table 10
	$_ir_{bis}$	Same as Table 10
	a_i	Same as Table 10
2. Item difficulty level	p_i	Same as Table 10
	$\bar{X}_i$	Same as Table 10
	P_i	Same as Table 10
	b_i	Same as Table 10
3. Relation of the item to test blueprint and/or domain specification	R_{kji}	Same as Table 10
	I_{ji}	Same as Table 10
	ID_{kli}	A code for the location of the item in a content by objectives grid (i.e., test blueprint)
4. Estimated total test statistics	$\tilde{\bar{X}}$	Estimated mean of the total test scores when the items selected so far are used.
	$\widetilde{SD}$	Estimated standard deviation of the total test scores when the items selected so far are used.
	$\widetilde{KR20}$	Estimated Kuder-Richardson test reliability when the items selected so far are used.

a program that would help teachers choose items. This program should use both a_i and b_i (i.e., the latent trait item difficulty index) to help a teacher design a test for measuring relative achievement using the item information function ($=\sum P(\Theta_a)_i$).

Our recommendations in this area are summarized in Table 21.

Insert Table 21 here

2. Item difficulty level. Item difficulty indices have been discussed previously and our recommendations for other purposes summarized in Table 3 and 14. Our recommendation for item difficulty indices in this section are the same as those for Table 14, except that we would recommend that the latent trait parameter b_i be incorporated into the item analysis program in the manner suggested above for using a_i, the latent trait item discrimination index.

Insert Table 22 here

3. Relation of item to test blueprint and/or domain specification. Our recommendations for these congruence indices are listed in Table 11 and as Equation [28] through [30]. We note here that in addition to a rating of how well a test item matches an objective, it is necessary to identify the content topic and level of understanding covered by each test item. This is not a statistic per se, but it is an index number that helps the teacher to identify the item and to check a test's balance of coverage.

Table 21. Recommended item statistics to use to help teachers select items to put on a classroom test: Item discrimination indices

Type of item scoring:	Basic: Should be included in every item analysis program, if at all possible. Routinely present to or interpret for a teacher on every test item.	Basic: Make available to teachers upon their request only.	Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
Dichotomous ($Y_{ai} = 0,1$)	D_i		a_i	${}_ir_{bis}$ ${}_ir_{ptbis}$
Graded or continuous ($m_i \le Y_{ai} \le 1_i$)	D_i'			${}_ir_{ptpolyserial}$ ${}_ir_{polyserial}$

Note: See the text for definitions, formulas, and explanations related to Equations [35] - [39].

Table 22. Recommended item statistics to use to help teachers select items to put on a classroom test: Item difficulty indices

Type of item scoring:	Basic: Should be included in every item analysis program, if at all possible.		Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
	Routinely present to or interpret for a teacher on every test item.	Make available to teachers upon their request only.		
Dichotomous ($Y_{ai} = 0,1$)	p_i		b_i	
Graded or continuous ($m_i \leq Y_{ai} \leq l_i$)	p_i	$\bar{Y}_i$		

Note: See the text for definitions, formulas, and explanations related to Equations [1] - [3] and [6] - [8].

We call this index number:

$$In_{kli} = \text{index number of the } \underline{i\text{th}} \text{ item in relation to the } \underline{l\text{th}} \text{ topic in the unit and the } \underline{k\text{th}} \text{ level of understanding} \quad [47]$$

Insert Table 23 here

4. Using combinations of indices to select items for classroom tests. The item statistics identified above cannot be used independently, but must be used in combination. The particular combinations to use depend on the type of decisions for which the test is to be used and, in particular, on whether the test is to be used to measure absolute or relative achievement and whether partial or complete ordering is desired. If latent trait parameters are available and the measurement of relative achievement is desired, then the microcomputer program can use a_i, b_i, and c_i in connection with the item information function to help design a test that will provide the most information possible at certain ability levels. Lord (1980) provides guidelines for this process.

But most teachers will not have access to items precalibrated by latent trait methods. Rather, they will have items for which are available simply item difficulty, item discrimination, and some indication of what the item is measuring. We recommend that the item analysis program incorporate some rules of thumb that will help the teacher to select items using the latter statistical indices when the test purpose is specified. Table 24 summarizes the rules of thumb we recommend. The rules of thumb in this table are consistent with modern concepts of item analysis and test design as these have le 23. Recommended item statistics to use to help teachers select items to put on a classroom test: Relation of the item to the test blueprint and/or domain specification

	Basic: Should be included in every item analysis program, if at all possible.		Recommended: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	Not recommended for item analysis programs serving the above mentioned purposes.
e of item ring:	Routinely present to or interpret for a teacher on every test item.	Make available to teachers upon their request only.		
hotomous $_i$ = 0,1)	R_{kji} ID_{kii}	Mdn_{Rji} I_{ji}		
ded or ntinuous $\le Y_{ai} \le 1_i$)	R_{kji} ID_{kii}	Mdn_{Rji} I_{ji}		

: See the text for definitions, formulas, and explanations related to Equations [28] - [30] and [47].

articulated by Lord (1953, 1980) and Henrysson (1971).

Insert Table 24 here

5. Estimated total test statistics. An item analysis program that is to be used to help teachers select items should provide estimates of the properties of the total test scores based on the selected items. The item statistics recommended in Table 21 and 22 for dichotomous items can be used to estimate the test mean, standard deviation, and Kuder-Richardson formula 20 reliability as follows:

$$_{I}\tilde{\bar{X}} = \sum_{i=1}^{I} p_i \qquad , \quad Y_{ai} = 0,\ 1 \qquad [48]$$

= estimated mean of the test composed of I items

$$_{I}\tilde{SD}_X = \frac{\sum_{i=1}^{I} D_i}{\sqrt{6}} \qquad , \quad Y_{ai} = 0,\ 1 \qquad [49]$$

$$_{I}\tilde{KR20} = \left[\frac{I}{I-1}\right] \left[- \frac{\sum_{i\neq 1}^{I} p_i(1-p_i)}{(\Sigma D_i)/\sqrt{6}} \right], \quad Y_{ai} = 0,\ 1 \qquad [50]$$

where

$i = 1, 2, \ldots, I$ indexes the items selected for the test

D_i = the net D discrimination index for dichotomously scored Item i

p_i = the difficulty index for dichotomously scored Item i

Table 24. Rules of thumb for using item analysis data to build classroom tests.

	Relative achievement is the focus		Absolute achievement is the focus
	Complete ordering	**Partial ordering** (two groups)	
General concerns	Ranking all the pupils in terms of their relative attainment in a subject area.	Dividing pupils into two groups on the basis of their relative attainment. Pupils within each group will be treated alike.	Assess the absolute status (achievement) of the pupil with respect to a well-defined domain of instructionally relevant tasks
Specific focus of test	Seek to accurately describe differences in relative achievement between individual pupils.	Seek to accurately classify persons into two categories.	Seek to accurately estimate the percentage of the domain each pupil can perform successfully.
Attention to the test's blueprint	Be sure that items cover all important topics and objectives within the blueprint.	Be sure that items cover all important topics and objectives within the blueprint.	Be sure items are a representative, random sample from the defined domain which the blueprint operationalizes.
How the difficulty index *(p)* is used	Within each topical area of the blueprint, select those items with: (1) *p* between 0.16 and 0.84 if performance on the test represents a single ability. (2) *p* between 0.40 and 0.60, if performance on the test represents several different abilities. Note. Items should be easier than described above if guessing is a factor.	Within each topical area of the blueprint, select those items with *p*-values slightly larger than the percentage of persons to be classified in the upper group [e.g., if the class is to be divided in half (0.50) then items with *p*-values of about 0.60 should be selected, if the division is lower 75% vs. upper 25%, items should have *p* = 0.35 (approximately)]. Note: The above suggestion assumes the test measures a single ability.	Don't select items on the basis of their *p*-values, but study each *p* to see if it is signaling a poorly written item
How the discrimination index *(D)* is used	Within each topical area of the blueprint, select items with *D* greater than or equal to +0.30.	Within each topical area of the blueprint, select items with *D* greater than or equal to +0.30.	All items should have *D* greater than or equal to 0.00. Unless there is a rational explanation to the contrary, revise those items not possessing this property.

Source: Nitko (1983, pg. 301)

Expression [49] was derived by Ebel (1967) under the assumption that the test scores are normally distributed. The sampling distribution and standard errors of these estimates are unknown and the effect of non-normality on equations [49] and [50] is unknown. Expression [48] does not depend on distribution assumptions.

If items are scored continuously, then [48] becomes

$$_2\tilde{\bar{X}} = \sum_{i=1}^{I} \bar{Y}_{\cdot i} \quad , \quad m_i \leq Y_{ai} \leq 1_i \tag{51}$$

The following expressions relate item scores (either continuous or dichotomous) to total test score standard deviation and reliability

$$_2\tilde{SD}_X = \sum_{i=1}^{I} (r_{Y_{ai}X_a}) SD_{Y_{ai}} \tag{52}$$

$$\tilde{\alpha} = \left(\frac{I}{I-1}\right)\left[1 - \frac{\sum_{i=1}^{I}(SD_{Y_{ai}})^2}{(_2\tilde{SD}_X)^2}\right] \tag{53}$$

In the above formulas $r_{Y_{ai}X_a}$ is the Pearson product moment correlation between the item scores and the total test score on the tryout edition of the test. If Y_{ai} is dichotomous, then this correlation becomes the $_ir_{ptbis}$. Equations [52] and [53] are useful and may provide better estimates than [49] and [50]. It is recommended that $_ir_{ptbis}$ be corrected so that it estimates the correlation of each item with the common true score measured by the whole set of items as suggested by Henrysson (1971).

If $r_{Y_{ai}X_i}$ is unknown, Thorndike (1982) suggests estimating its mean value, $\bar{r}_{Y_{ai}X_i}$, from past experience and substituting this estimate in [52] and [53]. Further, if the items are dichotomous and the average difficulty of the items on the test, $\bar{p}_i$, can be estimated equations [51]-[53] can be simplified as follows:

$$_3\tilde{\bar{X}} = I\bar{p} \qquad [54]$$

$$_3\widetilde{SD}_X = I\sqrt{\bar{p}(1-\bar{p})}\;\overline{r_{Y_{ai}X_i}} \qquad [55]$$

$$_3\widetilde{KR20} = \left(\frac{I}{I-1}\right)\left(1-\frac{1}{I(\bar{r}_{Y_{ai}X_i})^2}\right) \qquad [56]$$

It should be noted that [55] overestimates the standard deviation (Thorndike, 1982).

Summary. Our recommendations for this section are summarized in Table 25.

Insert Table 25 here

Empirical Data Concerning the Sampling Fluctuations in Selected Item Statistics

In an effort to obtain more information about the sampling fluctuations of some of the item statistics recommended in this report, we undertook a sampling study with the assistance of Dr. Huynh Huynh of the University of South Carolina. We sought to simulate the fluctuations in students that might occur from year to year in a teacher's class. To do this we used the item response data bank available at the University of South Carolina in connection with technical research conducted by the Mastery Testing Project (NIE-G-78-0087) and the Technical Works of Basic Skills Assessment Programs

Table 25. Recommended item statistics to help a teacher select items for a test: Estimating total test properties.

	<u>Basic</u>: Should be included in every item analysis program, if at all possible.		<u>Recommended</u>: Useful to include if (a) research shows teachers can use and (b) microcomputer has sufficient memory and speed.	<u>Not recommended</u> for item analysis programs serving the above mentioned purposes.
Type of item scoring:	Routinely present to or interpret for a teacher on every item selection situation.	Make available to teachers upon their request only.		
Dichotomous ($Y_{ai} = 0,1$)	$_3\tilde{\bar{X}}$ $_3\tilde{SD}_X$ $_3\tilde{KR20}$		Item information function	$_1\tilde{\bar{X}}$ $_1SD_X$ $_1\tilde{KR20}$
Graded or continuous ($m_i \leq Y_{ai} \leq 1_i$)	$_2\tilde{\bar{X}}$ $_2\tilde{SD}_X$ $\tilde{\alpha}$			

Note: See the text for definitions, formulas, and explanations related to formulas [48] - [56].

Project (NIE-G-80-0119) as these were applied to the South Carolina Basic Skills Assessment Program (SCBSAP). A basic description of the SCBSAP is given in Huynh and Castell (1982).

The data base used in our study consisted of responses from 2400 students in each of several grades who had taken the Mathematics and Reading tests of the SCBAP in 1981. This large group was selected as a stratfified cluster sample of the South Carolina student population. The Reading test contained 36 items and the Mathematics test contained 30 items. Within each grade level four items were selected for study. In the population of 2400 students the items selected had p-values between approximately 0.85 and 0.55, the range of p-values we believe is likely to be encountered in teacher-made domain-referenced tests.

To simulate fluctuations from sample to sample 80 random samples of 30 students each were selected and the various item statistics were computed for each sample. The samples were selected such a way that some (if not all) of the 30 students within a sample were from the same classroom. We note that the class-to-class or year-to-year fluctuations experienced by a teacher are likely to be less variable than fluctuations based on simple random sampling since a teacher will generally use a test either within the same school building (usually associated with a neighborhood) or in different buildings but within the same school district. Simple random samples from a state's population should be more variable since any one sample would contain students from widely scattered school districts with quite diverse characteristics.

It is likely, however, that the sampling distributions we report are more variable than a teacher might experience, lying somewhere between a distribution of strictly random samples and a distribution of within classroom samples over years. This is because, although we sampled students within a classroom, students in the subsequent sample came from another school distric .

In this paper we report only the preliminary results, since the study is on-going. We report sampling fluctuations for the following statistics: item difficulty, item discrimination, proportion in each third of the class, modified caution index for items, and chi- square. Each statistic is computed for each of four items as follows:

Reading		Mathematics	
Grade 1	Grade 6	Grade 2	Grade 6
Item 18	Item 34	Item 4	Item 21
$\phi = 0.597$	$\phi = 0.560$	$\phi = 0.564$	$\phi = 0.559$
$b = 0.959$	$b = 0.785$	$b = 2.288$	$b = -0.011$

Here, ϕ is the proportion of the 2400 students answering the item correctly and b is the Rasch item difficulty for three items. Because this is a preliminary report of our empirical study, we have not reported data on the other items investigated.

Table 26 shows the empirical sampling distribution of the item difficulty index, p_i, for each of the four items. The four distributions are roughly comparable. Sample p-values range from approximately .84 to .30. The mean of each distribution is reasonably close to its expected value, ϕ. However, the distributions are slightly more variable than expected. The standard error of a proportion based on random samples is

$$\sigma_p = \sqrt{\frac{\phi(1-\phi)}{N}}$$

where ϕ is the population proportion and N is the sample size. For each of the distributions in Table 26, σ_p is approximately 0.09, whereas the actual standard deviations are around 0.10.

Insert Table 26 here

Table 26. Empirical sampling distributions for the item difficulty index p

Values of p	Reading: Grade 1 Item 18 ϕ = 0.597 b = 0.959	Reading: Grade 6 Item 34 ϕ = 0.560 b = 0.785	Mathematics: Grade 2 Item 4 ϕ = 0.564 b = 2.288	Mathematics: Grade 6 Item 21 ϕ = 0.559 b = -0.011
.95 - 1.00				
.90 - .94				
.85 - .89				
.80 - .84	1	1		1
.75 - .79	4	1	1	3
.70 - .74	8	7	4	6
.65 - .69	8	6	5	4
.60 - .64	23	20	18	19
.55 - .59	9	9	10	8
.50 - .54	15	15	17	14
.45 - .49	10	7	9	11
.40 - .44	1	7	9	10
.35 - .39	1	3		2
.30 - .34		4	1	2
.25 - .29			1	
.20 - .24				
.15 - .19				
.10 - .14				
.05 - .09				
.00 - .04				
Mean	.59	.55	.56	.55
Std. Dev.	.09	.11	.10	.11
No. of samples	80	80	80	80

Table 27 summarizes the empirical distributions of several item discrimination indices. The distributions behave as expected. Note that the net D index was computed on the basis of upper and lower thirds and upper and lower halves. As expected the items show less discrimination when the halves are used compared to the thirds: The mean discrimination index for the halves' distributions run approximately 0.10 to 0.14 lower than the means of the thirds distributions. Since on the average the persons in the halves groups are closer in ability to each other than are the average persons in the thirds group, this result is expected. Further, since there are more students in the halves groups than in the thirds groups (15 vs. 10 students) the sampling distribution of net D when computed on halves is less variable.

With a lower mean discrimination value and less variability, more poorly discriminating items would be identified if the upper and lower groups consisted of the halves of the class rather than the upper and lower thirds. For example, if $D_i < 0.30$ is used as a rule of thumb for flagging a poorly discriminating item, then in 80 replications, Item 18 would be flagged 1 time using the thirds procedure vs. 8 times with the halves procedure, Item 34 one time vs. 10 times, Item 4 five times ve 26 times, and Item 21 twelve times ve 26 times. We would take a conservative view stating that it is better to flag an item and have a teacher check it than to let the item go by unreviewed. Thus, we would recommend using the upper and lower halves for the net D index.

Insert Table 27 here

Table 28 shows distribution of the proportions of students passing an item in the upper, middle, and lower thirds of the class based on the total

Table 27. Empirical sampling distribution for item discrimination indices (D 1/3 = net D computed using upper and lower thirds of the class, D 1/2 = net D computed using the upper and lower halves, BIS = biserial correlation, P-BIS = point biserial correlation).

	READING								MATH							
	GRADE 1, ITEM 18 (p = 0.597)				GRADE 6, ITEM 34 (p = 0.560)				GRADE 2, ITEM 4 (p = 0.564)				GRADE 6, ITEM 2 (p = 0.599)			
	D 1/3	D 1/2	BIS	P-BIS	D 1/3	D 1/2	BIS	P-BIS	D 1/3	D 1/2	BIS	P-BIS	D 1/3	D 1/2	BIS	P-BIS
.95 - 1.00+	1		11				6									
.90 - .94	10		4		9		9		3		1				1	
.85 - .89		1	8				8				2				1	
.80 - .84	9		4	2	12	1	10	1	9		5		3		3	
.75 - .79		2	4	6			7	4			7				4	
.70 - .74	14		8	5	17	3	8	11	8	3	7	2	9	1	9	1
.65 - .69		14	9	13		10	6	12		4	6	2		1	8	3
.60 - .64	14	10	11	3	22	10	10	11	15	5	10	8	18	4	8	3
.55 - .59			4	11			6	9		0	7	9			8	10
.50 - .54	17	12	5	11	8	12	2	9	19	10	6	10	15	7	9	12
.45 - .49		10	4	10		17	4	10		8	8	10		13	4	11
.40 - .44	10	11	2	5	6	13	1	5	13	14	7	9	11	20	11	8
.35 - .39			4	6				4		0	6	9			2	8
.30 - .34	4	12	1	5	5	4	2	1	8	10	2	12	12	8	2	12
.25 - .29		5	1	1		4	1			10	2	2		9	4	
.20 - .24		3		1	1	4		3	1	6	1	3	8	9	2	5
.15 - .19				1						0	1	1			4	3
.10 - .14	1					1			2	5	0	1	3	2		1
.05 - .09						1				3	2	2		3		3
.00 - .04										1				2		
(-.05) - (-.01)																
(-.10) - (-.06)									2	1			1	1		
Mean	.62	.48	.70	.55	.64	.48	.73	.57	.52	.38	.56	.44	.46	.36	.54	.41
std. dev.	.19	.16	.19	.15	.17	.15	.17	.13	.20	.18	.19	.15	.19	.16	.19	.16
Number of Samples	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80	80

test score. The sampling distributions are as expected: lower third students answering the item correctly in fewer numbers than the middle and upper third and variability as indicated by sampling theory. An exception to this statement is the middle third of the students on Item 34. This group seems to be more variable than expected. It appears that some useful information for teachers can be obtained by displaying these proportions for each item in each class.

Insert Table 28 here

Table 29 shows the sampling distributions of the modified caution index for items. This index is designed to identify items exhibiting unusual responses compared to the other items in the test. Since the four items in Table 28 are part of a large scale testing program in which the items were professionally review, tried-out, and selected, we would not expect high values of this caution index in Table 29. This appears to be upheld. Virtually all of the values of the caution index are below 0.55. Thus, none of these items would likely have been brought to a teacher's attention as unusual in their performance relative to other items in the test. We recommend that this index be incorporated into the instructional improvement and guidance section of an item analysis program if a microcomputer can handle it.

Insert Table 29 here

Table 30 shows the distributions of the frequency chi-squares, SSQ_i, which test whether the upper scoring group follow a guessing pattern (i.e., a uniform distribution). We expect that with SCBSAP items, upper group students would not guess. Thus, SSQ_i-values should be large and the hypotheses of

Table 28. Empirical sampling distributions of the proportion of each third of a sample answering an item correctly.

	Reading						Mathematics					
	Grade 1			Grade 6			Grade 2			Grade 6		
	Item 18 (p = 0.597)			Item 34 (p = 0.560)			Item 4 (p = 0.564)			Item 21 (p = 0.599)		
	lower 1/3	middle 1/3	upper 1/3	lower 1/3	middle 1/3	upper 1/3	lower 1/3	middle 1/3	upper 1/3	lower 1/3	middle 1/3	upper 1/3
.95 - 1.00		1	38		2	21		1	14			11
.90 - .94		2	27		3	28		2	27		1	25
.85 - .89												
.80 - .84		7	10	1	12	23		10	19	1	6	17
.75 - .79												
.70 - .74	1	13	5		13	5	1	10	10	3	7	12
.65 - .69												
.60 - .64	5	15			9	2	3	11	5	5	21	9
.55 - .59												
.50 - .54	10	19		2	11	1	11	26	3	9	22	5
.45 - .49												
.40 - .44	15	12		15	18		13	13	2	15	9	1
.35 - .39												
.30 - .34	19	8		17	7		22	2		24	11	
.25 - .29												
.20 - .24	14	3		20	2		15	5		12	3	
.15 - .19												
.10 - .14	14			20	2		12			10		
.05 - .09												
.00 - .04	2			5	1		3			1		
Mean	.31	.55	.92	.24	.55	.87	.30	.55	.82	.34	.52	.80
Std. Dev.	.16	.17	.09	.14	.22	.11	.16	.18	.15	.17	.16	.15
No. of Samples	80	80	80	80	80	80	80	80	80	80	80	80

Table 29. Empirical sampling distributions for the modified caution index for items.

	Reading		Mathematics	
	Grade 1	Grade 6	Grade 2	Grade 6
	Item 18 (p = 0.597)	Item 34 (p = 0.560)	Item 4 (p = 0.564)	Item 21 (p = 0.599)
.95 - 1.00				
.90 - .94				
.85 - .89				
.80 - .84				
.75 - .79			1	
.70 - .74				
.65 - .69				
.60 - .64		1	2	
.55 - .59	1		1	
.50 - .54	1	1	4	1
.45 - .49	1	1	4	1
.40 - .44	1	5	6	3
.35 - .39	1	1	8	5
.30 - .34	6	3	7	8
.25 - .29	13	12	9	6
.20 - .24	16	14	10	20
.15 - .19	13	13	11	15
.10 - .14	13	18	9	14
.05 - .09	12	5	5	6
.00 - .04	2	6	3	1
Mean	.20	.20	.28	.22
Std. Dev.	.11	.12	.15	.10
No. of Samples	80	80	80	80

of a uniform distribution would be rejected. Table 30 shows that the rate retention of the hypotheses of a uniform distribution is quite small. (Items from grades 1 and 2 have 3 alternatives and items from grade 6 have 4 alternatives. Thus, the degrees of freedom are 3 and 4, respectively.) Thus, from this preliminary data our original fear of a large Type II error rate is not upheld.

Insert Table 30 here

SUMMARY

We have reviewed fifty or so statistics in this report in relation to their usefulness for an item analysis microcomputer program that is intended to be appropriate for the analysis of domain-referenced classroom tests. We took the view that the primary purposes of an item analysis of classroom tests are to: (a) inform the teacher about the strengths and weaknesses of the class in relation to the skills measured by the individual test items and (b) inform the teacher about the items that do not seem to be functioning well so that the teacher can rewrite or otherwise revise these items. A secondary purpose of a classroom item analysis program is to select items from a pool of items (an item bank) to put on a particular test in order to improve the utility of that test for a particular purpose.

In order to provide a context in which to review item statistics we define three broad areas of information a teacher would need in relation to test items. Then we specified the particular information needs which item-based information can serve under each of these three broad areas and how these particular kinds of information can link together testing and instruction.

Table 30. Empirical sampling distribution of the chi-square statistic SSQ_i for testing whether students in the upper group responded randomly to the items

	Reading		Mathematics	
	Grade 1	Grade 6	Grade 2	Grade 6
	Item 18	Item 34	Item 4	Item 21
95 - 100+				
90 - 94				
85 - 89				
80 - 84				
75 - 79				
70 - 74				
65 - 69				
60 - 64				
55 - 59				
50 - 54	8	3	1	1
45 - 49				
40 - 44	6	10	4	5
35 - 39	18	12	11	10
30 - 34	10	2	7	5
25 - 29	16	12	12	10
20 - 24	13	21	15	10
15 - 19	8	12	16	25
10 - 14	1	3	11	10
5 - 9		5	3	4
0 - 4				
Mean	31.57	27.81	24.49	24.02
Std. Dev.	9.59	10.73	9.78	9.77
N	80	80	80	80

Next we considered in relation to each specific type of information several statistical indices which seemed to provide the information required. We reviewed each statistic in terms of its statistical and numerical properties, its suitability for the type of data likely to be encountered with classroom tests, its ability to be understood by teachers, and the practicality of computing it on a microcomputer of the type typically found in schools. As a result of this analysis, we prepared to each specific type of information our recommendations in relation to each statistic. For each type of information we classified the statistics reviews as (a) basic (to be included in every item analysis program if at all possible), (b) recommended (useful statistics that should be included if the microcomputer has sufficient memory and speed and if research shows that teachers can use them) and (c) not recommended (for item analysis microcomputer programs that are intended to serve the purposes we outlined).

In addition to this literature review, we reported some preliminary results of an empirical sampling study we are in the process of undertaking to study the sampling fluctuations of some of the recommended item analysis statistics. The preliminary results of this empirical study indicated that the recommendations we made we generally upheld by data from classrooms. Further, the empirical results offered guidelines for setting rules of thumb for the numerical value of statistics to use when flagging an item and bringing it to the attention of the teacher.

References

Airasian, P. W., & Bart, W. M. Ordering Theory: A new and useful measurement model. Journal of Educational Technology, May 1973, 56-60.

Bart, W. M., & Krus, D. J. An ordering-theoretic method to determine hierarchies among items. Educational and Psychological Measurement, 1973, 33, 291-300.

Berk, R. A. Item analysis. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980.

Bormuth, J. R. Development of readability analyses. (Final Report, Project No. 7-0052, Contract No. OEG-3-7-070052-0326) Washington, D.C.: Office of Education, Bureau of Research, U. S. Department of Health, Education, and Welfare, March 1969.

Brennan, R. L. A generalized upper-lower item discrimination index. Educational and Psychological Measurement, 1972, 32, 289-303.

Brennan, R. L. The evaluation of mastery test items. (Final Report, Project No. 2B118) Washington, D.C.: National Center for Educational Research and Development, U. S. Office of Education, 1974.

Brennan, R. L. Applications of generalizability theory. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980.

Carroll, J., Davies, P., & Richman, B. Word frequency book. New York: American Heritage Publishing Co., 1971.

Cox, R. C., & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. A paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1966.

Dale, E., & Eichholz, G. Children's knowledge of words: An interim report. Columbus, OH: Bureau of Educational Research and Service, The Ohio State University, 1980.

Dale, E., & O'Rourk, J. The living word vocabulary--The words we know. Elgin, IL: Dome Press, Inc., 1976.

Darlington, R. B., & Bishop, C. H. Increasing test validity by considering inter-item correlations. Journal of Applied Psychology, 1966, 50, 322-330.

Donlon, T. F., & Fischer, F. E. An index of an individual's agreement with group-determined item difficulties. Educational and Psychological Measurement, 1968, 28, 105-113.

Ebel, R. L. The relation of item discrimination to test reliability. Journal of Educational Measurement, 1967, 4, 125-128.

Ebel, R. L. Essentials of educational measurement (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall, 1972.

Ebel, R. L. Essentials of educational measurement (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc., 1979.

Findley, W. G. A rationale for the evaluation of item discrimination statistics. Educational and Psychological Measurement, 1956, 16, 175-180.

Fry, E. Extended graph and rules for readability. Publisher's Weekly, October 29, 1979, p. 41.

Gagné, R. M. Learning hierarchies. Educational Psychologist, 1968, 6, 1-9.

Greeno, J. G. Cognitive objectives of instruction: Theory of knowledge for solving problems and answering questions. In D. Klahr (Ed.,), Cognition and instruction. Hillsdale, NJ: John Wiley & Sons, 1976.

Gregg, L. W. Methods and models for task analysis in instructional design. In D. Klahr (Ed.), Cognition and instruction. Hillsdale, NJ: John Wiley & Sons, 1976.

Guttman, L. Relation of scalogram analysis to other techniques. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clauser (Eds.), Studies in social psychology in World War II: Measurement and prediction (Vol. 4). Princeton, NJ: Princeton University Press, 1950.

Hambleton, R. K. Test score validity and standard-setting methods. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: Johns Hopkins University Press, 1980.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Harnisch, D. L., & Linn, R. L. Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 1981, 18, 133-146.

Henryssen, S. Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Hsu, T. C. Empirical data on criterion-referenced tests. A paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Hsu, T. C., & Khampalikit, C. Item analysis procedures for detecting group differences in guessing. Pittsburgh, PA: Programs in Educational Research Methodology, University of Pittsburgh, 1980.

Hsu, T. C., & Liou, C. S. Blind guessed multiple-choice items identified by examinees. A paper presented at the annual meeting of the National Council on Measurement in Education, New York, 1982.

Huynh, H. Item and test statistics for teacher-made tests. Unpublished memo, February 18, 1983.

Huynh, H., & Casteel, J. Technical works of basic skills assessment programs. (Final Report) Columbia, SC: College of Education, University of South Carolina, 1982.

Huynh, H., & Saunders, J. C. Solutions for some technical problems in domain-referenced mastery testing. (Final Report) Columbia, SC: College of Education, University of South Carolina, 1980.

Instructional Objectives Exchange. South Carolina Word List: Grades 1-12. Los Angeles, CA: Author, 1980.

Johnson, A. P. Notes on a suggested index of item validity: The U-L index. Journal of Educational Psychology, 1951, 42, 499-504.

Khampalikit, C. Race and sex differences in guessing behavior on a standardized achievement test in the elementary grades. Unpublished doctoral dissertation, School of Education, University of Pittsburgh, Pittsburgh, PA, 1982.

Klein, S. P., & Kosecoff, J. B. Issues and procedures in the development of criterion-referenced tests. In W. A. Mehrens (Ed.), Readings in measurement and evaluation in education and psychology. New York: Holt, Rinehart, and Winston, 276-293.

Kosecoff, J. B., & Klein, S. P. Instructional sensitivity statistics appropriate for objectives-based test items. (CSE Report No. 91). Los Angeles: Center of Study of Education, University of California, 1974.

Lord, F. M. The relation of test scores to the trait underlying the test. Educational and Psychological Measurement, 1953, 13, 517-549.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman (Ed.), Computer-assisted instruction, testing, and guidance. New York: Harper and Row, Publishers, 1970.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, 1980.

Mager, R. F. Measuring instructional intent: Or got a match? Belmont, CA: Lear Siegler, Inc./Fearon Publishers, 1973.

Nitko, A. J. Problems in the development of criterion-referenced tests: The IPI Pittsburgh experience. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. (CSE Monograph Series in Evaluation No. 3), Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Nitko, A. J. Distinguishing the many varieties of criterion-referenced tests. Review of Educational Research, 1980, 50, 461-485.

Nitko, A. J. Educational tests and measurements: An introduction. San Diego: Harcourt Brace Jovanovich, 1983.

Olsson, U., Drasgow, F., & Dorans, N. J. The polyserial correlation coefficient. Psychometrika, 1982, 47, 337-347.

Pike, L. W. Implicit guessing strategies of GRE-Aptitude examinees classified by ethnic group and sex. (ETS GRE 75-10) Princeton NJ: Educational Testing Service, 1980.

Pike, L. W., & Flaugher, R. L. Assessing the meaningfulness of group responses to multiple-choice test items. Proceedings of the 78th Annual Convention of the American Psychological Association, 1970, 101-102.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Rasch, G. Probabilistic models for some intelligence and educational tests. Copenhagen, Denmark: The Danish Institute for Educational Research, 1960.

Resnick, L. B. Task analysis in instructional design: Some cases from mathematics. In D. Klahr (Ed.), Cognition and instruction. Hillsdale, NJ: John Wiley & Sons, 1976.

Roudabush, G. E. Item selection for criterion-referenced tests. A paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973.

Rovinelli, R. J., & Hambleton, R. K. On the use of content specialists in the assessment of criterion-referenced test item validity. Dutch Journal of Educational Research, 1977, 2, 49-60.

Rudner, L. M. Individual assessment accuracy. Unpublished paper. Washington, D.C.: National Institute of Education, 1982.

Sakiey, E., & Fry, E. 3,000 instant words. N.J.: Dreir Educational Systems, 1979.

Sax, G. Principles of educational and psychological measurement and evaluation (2nd ed.). Belmont, CA: Wadsworth Publishing Co., 1980.

Takeya, M. A study on item relational structure analysis of criterion-referenced tests. Unpublished doctoral dissertation. Tokyo, Waseda University, June, 1981. (Cited in Tatsuoka & Tatsuoka, 1981).

Tatsuoka, K. K. An approach to assessing the seriousness of error types and predictability of future performance. (Research Report 81-1) Urbana, IL: Computer-based Education Research Laboratory, University of Illinois, 1981.

Tatsuoka, K. K., & Tatsuoka, M. M. Item analysis of tests designed for diagnosing bugs: Item relational structure analysis method. (Research Report 81-7) Urbana, IL: Computer-based Education Research Laboratory, University of Illinois, 1981.

Tatsuoka, K. K., & Tatsuoka, M. M. Detection of aberrant response patterns and their effect on dimensionality. Journal of Educational Statistics, 1982, 7, 215-231.

Taylor, S., et al. EDL core vocabularies in reading, mathematics, sciences, and social studies. New York: McGraw-Hill, 1979.

Thorndike, R. L. Applied psychometrics. Boston: Houghton Mifflin Co., 1982.

White, G. W., Feldt, L. S., & Sabers, D. L. Relative effectiveness of three item selection procedures for maximizing test reliability. A paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C., April, 1975.

Whitney, D. R., & Sabers, D. L. Improving essay examinations III: Use of item analysis. (Technical Bulletin No. 11) Iowa City, IA: University Evaluation and Examination Service, The University of Iowa, May, 1970.

Wise, S. L. A modified order-analysis procedure for determining unidimensional items sets. Unpublished doctoral dissertation, University of Illinois, 1981.

Wright, B. D., & Stone, M. H. Best test design. Chicago: MESA Press, 1979.